THESIS

EXTREME PRECIPITATION AND FLOODING: EXPOSURE CHARACTERIZATION AND THE ASSOCIATION BETWEEN EXPOSURE AND MORTALITY IN 108 UNITED STATES COMMUNITIES, 1987–2005

Submitted by

Rachel Severson

Department of Environmental and Radiological Health Sciences

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2017

Master's Committee:

Advisor: Brooke Anderson

Jennifer Peel
Neil Grigg

ABSTRACT

EXTREME PRECIPITATION AND FLOODING: EXPOSURE CHARACTERIZATION AND THE ASSOCIATION BETWEEN EXPOSURE AND MORTALITY IN 108 UNITED STATES COMMUNITIES, 1987-2005

There is substantial evidence that extreme precipitation and flooding are serious threats to public health and safety. These threats are predicted to increase with climate change. Epidemiological studies investigating the health effects of these events vary in the methods used to characterize exposure. Here, we compare two sources of precipitation data (National Oceanic and Atmospheric Administration (NOAA) monitor-based and North American Land Data Assimilation Systems (NLDAS-2) Reanalysis data-based) for estimating exposure to extreme precipitation and two sources of flooding data, based on United States Geological Survey (USGS) streamflow gages and the NOAA Storm Events database. We investigate associations between each of the four exposure metrics and short-term risk of four causes of mortality (accidental, respiratory-related, cardiovascular-related, and all-cause) in the U.S. from 1987 through 2005. Average daily precipitation values from the two precipitation data sources were moderately- to well-correlated (rho = 0.74); however, values from the two data sources were less correlated when comparing binary metrics of exposure to extreme precipitation days (J = 0.35). Binary metrics of daily flood exposure were generally poorly correlated between the two flood data sources (rho = 0.07; J = 0.05). There was generally little correlation between extreme precipitation exposure and flood exposure in study communities. We did not observe evidence of a positive association between any of the four exposure metrics and risk of any of the four mortality outcomes considered. Our results suggest, due to the observed lack of agreement

between different extreme precipitation and flood metrics, that exposure to extreme precipitation might not serve as an effective surrogate for exposures related to flooding. Furthermore, it is possible that extreme precipitation and flood exposures may often be too localized to allow accurate exposure assessment at the community level for epidemiological studies.

ACKNOWLEDGEMENTS

Thank you so much to my advisor, Dr. Brooke Anderson. Dr. Anderson has been an invaluable mentor and source of support throughout my time here at CSU. Thanks to Dr. Anderson I've been able to grow from a complete novice at coding to being able to co-write an R software package, and conduct the analyses for this project. I have been very grateful for the opportunity to work with Dr. Anderson—everything I've learned under her guidance will be crucial for my future endeavors. Thank you to Dr. Jennifer Peel for welcoming me into the Environmental Health program, and for guiding me through the program's requirements and through thinking about my future plans. Dr. Peel's classes have served as the foundation of my epidemiological education, and I'm grateful for her influence on this project. My outside committee member, Dr. Neil Grigg, has provided an important perspective throughout. I would like to thank him for his interest in my project, and for his suggestions regarding this project and future research. Thank you to my friends and family for their support throughout my time working through coursework and research. I would not have been able to complete either aspect of my degree without their encouragement. Finally, I would like to acknowledge funding from the National Institute of Health Sciences and the Colorado State University Water Center.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The patterns, intensities, and durations of extreme weather and climate events in the United States have changed over the past decades and are expected to continue to vary in the coming century with climate change. The health impacts of extreme precipitation and flooding in particular are multi-faceted and expected to worsen (Bell et al. 2016). There are several health outcomes of concern associated with these events, and several ways to measure their occurrence—this is exemplified by the breadth of health outcomes investigated using varying methods of measuring exposure in epidemiological studies in the existing literature.

Epidemiological evidence of health risks associated with extreme precipitation and flooding

Epidemiological research has identified a number of health risks potentially associated with extreme precipitation and flood events. For extreme precipitation, epidemiological studies have reported evidence of increased risk of gastrointestinal (GI) illness (Thomas et al. 2006; Colford et al. 1999; Curriero et al. 2001; Tornevi, Axelsson, and Forsberg 2013; Nichols et al. 2009), vehicular accident-related injuries and fatalities (Ashley et al. 2015; Black, Villarini, and Mote 2017), and respiratory-related health outcomes (Fisman et al. 2005; Dunn et al. 2012; Halsby et al. 2014; Brandsema et al. 2014; Soneja et al. 2016; Solomon et al. 2006).

Several studies have investigated whether rates of GI illness are associated with extreme precipitation. In a national study of United States (U.S.) communities, Curriero and coauthors analyzed waterborne disease outbreaks and total monthly precipitation in the United States from 1948 through 1994 and found that over half of GI outbreaks were preceded in the previous two

months by precipitation events above the 90th percentile of a community's typical precipitation, and 68% of events were preceded by events above the 80th percentile (Curriero et al. 2001). Thomas et al. (2006) conducted another spatially and temporally large-scale study, investigating associations between extreme precipitation and waterborne disease outbreaks in Canada between 1975 and 2001. They reported that extreme precipitation events were associated with an increase in the relative odds of an outbreak of 2.28 (95% confidence interval: 1.22, 4.29) (Thomas et al. 2006). Drayna et al. (2010) investigated a link between precipitation and acute gastrointestinal illness pediatric emergency department visits to the Children's Hospital of Wisconsin. The authors found an association between any rainfall 4 days prior and an 11% increase in acute gastrointestinal illness pediatric emergency department visits (Drayna et al. 2010). Conversely, Colford and coauthors (1999) investigated extreme precipitation and GI outcomes failed to identify an association. The authors investigated the association between rainfall and the frequency of sick leave use, meant to act as a surrogate for gastrointestinal illness among 449 U.S. Postal Service workers in Sacramento, CA for seven months between October 1, 1992 and April 30, 1993—no meaningful difference was reported in the incidence rates of sick leave use between days with and without precipitation (Colford et al. 1999). While these studies investigated non-fatal outcomes, for more vulnerable populations such as young children and the elderly, these gastrointestinal-related health outcomes could also increase risk of mortality (Lane et al. 2013).

One potential pathway between extreme precipitation events and gastrointestinal illness involves exposure to water from combined sewer and storm water systems—these systems may be overwhelmed during extreme events (Colford et al. 1999; Jagai et al. 2015). The outflows from these systems could contaminate surface or groundwater sources, and could release sewage

into local waterways, resulting in an increased exposure to waterborne pathogens (Drayna et al. 2010). Researchers have detected associations between heavy rainfall and the rate of emergency room visit for gastrointestinal illness in regions where combined sewer overflow outfalls to drinking water sources. These associations were especially strong among those 65 and older, although not statistically significant (Jagai et al. 2015). According to the United States Environmental Protection Agency, combined sewer and storm water systems serve about 860 communities in the United States, totaling about 40 million people (EPA, 2016). Most of these communities have populations under 10,000 people (cities like New York, Philadelphia, and Atlanta are notable exceptions) and are located in the Northeast U.S. and near the Great Lakes (EPA, 2016).

Another relevant pathway between extreme precipitation and health risks involves driving; the visibility impairments caused by heavy rainfall have been found to result in increased accident-related vehicular fatalities (Ashley et al. 2015; Black, Villarini, and Mote 2017; Bergel-Hayat et al. 2013). Ashley et al. (2015) analyzed data from 1994 through 2011 from the National Highway Traffic Safety Administration (NHTSA)'s Fatality Analysis Reporting System (FARS) to characterize the role of weather in fatal motor vehicle crashes (2015). The authors observed that of fatal crashes that occurred in adverse weather conditions, almost half (46%) occurred in rain. These crashes comprise 8% of all fatal crashes (Ashley et al. 2015). Ashley et al. argue that these counts, combined with other fog, smoke, or dust-related crashes, suggest that crashes due to weather-related obscured vision are an important threat to drivers, and the increased risk due to these conditions deserves research attention comparable to more high profile hazards like tornadoes and lightning (Ashley et al. 2015). In a matched pair analysis, in which each day with measurable precipitation was paired with a control day without

precipitation exactly one week before or after, Black and coauthors observed statistically significant increases in vehicle crashes (10% increase; 95% confidence interval: 9.8%, 10.2%) and injury rates (8% increase; 95% confidence interval: 7.5%, 8.1%) during rainfall days compared to dry days in six U.S. states (Black et al. 2017). The risk of crashes was found to increase with increasing daily rainfall totals (Black et al. 2017). Positive correlations have been observed between rainfall per month and the number of monthly vehicle accidents in other countries as well, such as France and the Netherlands (Bergel-Hayat et al. 2013).

In addition to gastrointestinal illness and accidental fatalities, heavy precipitation could result in higher risk of respiratory-related outcomes. For example, *Legionella* is an environmentally ubiquitous bacterium that has been implicated in a type of severe pneumonia called legionellosis, or Legionnaires' disease. Legionnaires' disease is relatively uncommon but still results in a substantial number of outbreaks and fatal cases, particularly among the elderly (Phin et al. 2014.; Farnham et al. 2014; Campese et al. 2011). A plausible pathway for the transmission of *Legionella* to humans involves aerosolization of the bacterium. The environmental conditions that encourage this process and the survival of the bacterium are poorly understood (Dunn et al. 2012). Many researchers have hypothesized that weather conditions, particularly temperature, relative humidity, and precipitation, could play a role in *Legionella* transmission and subsequent Legionnaires' disease case incidence (Falkinham et al. 2015; Halsby et al. 2014; Chen et al. 2014; Dunn et al. 2012; Brandsema et al. 2014; Hicks et al. 2007; Farnham et al. 2014). Associations have been detected between rainfall and Legionnaires' disease cases and outbreaks. Brandsema and coauthors investigated sporadic cases in the Netherlands from 2003 to 2011, and reported that long-lasting and intense rainfall, in addition to temperature, contributes to an increased Legionnaires' disease incidence. Fishman et al. (2005)

reported a positive exposure-response association between rainfall, relative humidity and Legionnaires' in Philadelphia, PA between 1995 and 2003. In a study of cases in England and Wales from 1993 through 2008, Halsby et al. (2014) found that there may be an association between temperature and rainfall and the risk of sporadic Legionnaires' disease. Not all studies investigating the association between Legionnaires' and weather have observed an effect; for example, Dunn et al. (2012) reported no statistically significant associations with Legionnaires' disease incidence and weather in Glasgow, United Kingdom after adjusting for year-by-year and seasonal variation in cases.

In addition to the potentially fatal pneumonia caused by exposure to *Legionella*, there is evidence that extreme precipitation events are associated with increased risk of hospitalization for asthma (Soneja et al. 2016). Soneja and coauthors reported that the observed association was particularly strong in summer months, and among youth and adults—weaker associations were observed among those 65 and older (Soneja et al. 2016). In related research, there is evidence of associations between acute asthma outbreaks and thunderstorms, potentially partially attributable to the increase in airborne concentrations pollen and fungal spores that occurs during thunderstorm events (Dabrera et al. 2013; W. Anderson et al. 2001; D'Amato, Liccardi, and Frenguelli 2007).

Depending on soil moisture conditions and land use or cover conditions, extreme precipitation can lead to flooding (Rowe and Villarini 2013). Floods are among the most dangerous natural disasters due to the number of people affected and due to the average mortality per flood event; this is especially true for flash floods (Lowe, Ebi, and Forsberg 2013). There are many pathways that result in injury or death during a flood or in its aftermath—older age increases the risk of death in all of these pathways (Lane et al. 2013). Epidemiological studies

have found evidence of associations between flooding and myriad health outcomes, including gastrointestinal illness (T. J. Wade 2004; Wade et al. 2014; Lin, Wade, and Hilborn 2015; Setzer and Domino 2004; Tak et al. 2007; Thomas et al. 2006; Ding et al. 2013; Cann et al. 2013), accidental injury and fatality (Sharif et al. 2012; Alderman, Turner, and Tong 2012; Kellar and Schmidlin 2012), pulmonary impacts (Robinson et al. 2011; Solomon et al. 2006), and cardiovascular-related disease (Vanasse et al. 2016).

As with extreme precipitation events, a well-studied health risk from flood exposure is risk of illness resulting from exposure to waterborne pathogens, given that the conditions caused by flooding are amenable to the spread of waterborne disease (Setzer and Domino 2004; Wade et al. 2004; Wade et al. 2014; Lin, Wade, and Hilborn 2015; Ahern et al. 2005; Cann et al. 2013). *Cryptosporidium*, *G. lamblia*, and *T. gondii* have been implicated in waterborne disease outbreaks in the United States (Setzer and Domino 2004). Little research has investigated linkages between flooding and *H. pylori*, *M. avium*; however, there is a potential for exposure. *M. avium* and adenoviruses are notable deviations from the theme of gastrointestinal-related illness: they both cause upper respiratory tract infections (Setzer and Domino 2004).

Wade and coauthors and Setzer and Domino both investigated the health impacts of a single flooding event: severe flooding in the Midwestern United States in the spring of 2001 and Hurricane Floyd's landfall in North Carolina in September of 1999, respectively (Wade et al. 2004; Setzer and Domino 2004). Wade and coauthors identified a decrease in the quality of source water during the flood period (April 14, 2001 and May 30, 2001)—concentrations of *Giardia* cysts and male-specific coliphages both increased (Wade et al. 2004). They conducted a survey to determine if the rates of gastrointestinal symptoms increased in association with the flood period or contact with floodwater and found that during the flood period, rates of

gastrointestinal symptoms were statistically significantly elevated (Wade et al. 2004). Further, they found that among participants aged 50 or older, rates of gastrointestinal symptoms were more elevated compared to the entire study population—this effect was particularly evident for severe diarrhea. When considering contact with flood water as the exposure of interest, self-reported flooding of the house or yard was strongly associated with gastrointestinal symptoms (Wade et al. 2004).

Setzer and Domino (2004), conversely, found little evidence of elevated waterborne disease risks following a severe flood. They evaluated Medicaid outpatient utilization related to six waterborne pathogens (*Cryptosporidium*, *Giardia lamblia*, *Toxoplasma gondii*, *Helicobacter pylori*, *Mycobacterium avium*, and adenoviruses) and compared Medicaid utilization for the associated waterborne diseases during for the pre-Hurricane Floyd to post-Floyd periods. They found no clear increase in visits related to the six selected pathogens, concluding that it is unclear whether an increased risk of exposure to waterborne pathogens resulted in increased use of the health care system following this extreme flooding event (Setzer and Domino 2004).

In a review of the health impacts of flooding worldwide, Ahern et al. (2005) drew inferences from 212 studies and reported that there is inconclusive evidence for increased rates of diarrheal deaths associated with flooding but there is evidence for increased risk of non-fatal outcomes following flooding, including from increased transmission of fecal-oral, vector-borne, and rodent-borne disease (Ahern et al. 2005). Wade et al. (2014) and Lin et al. (2015) both conducted longer-term investigations of the association between gastrointestinal illness and flooding in Massachusetts (2003 through 2007 and 2003 through 2009, respectively). Wade et al. (2014) found that 7% of Emergency Room gastrointestinal-related visits were associated with flood events. Lin and coauthors found that in the 7 to 13 days following a flood, there was an

elevated rate of Emergency Room and outpatient visits for *Clostridium difficile* infection—*C. difficile* is a water-borne bacterium, and the primary cause of hospital-acquired infectious diarrhea in the United States (Lin et al. 2014).

Ding et al. (2013) found that floods increase the risk of diarrhea in Chinese people living along the Huaihe River, and that long-term, moderate floods may be more concerning regarding disease burden compared to shorter, more severe floods. In a systematic review of 83 papers investigating associations between waterborne disease outbreaks and extreme water-related weather events (the majority of which were in North America) Cann et al. (2011) found that heavy rainfall and flooding commonly preceded waterborne disease outbreaks. A possible pathway for these outbreaks is the contamination of the of drinking water supply (Cann et al. 2011).

In addition to the concern about disease caused by exposure to waterborne pathogens resulting from flood events, there is concern about event-related, accidental fatalities related to the event or the restoration process (Hajat et al. 2005; Kellar and Schmidlin 2012; Alderman et al. 2011; Ahern et al. 2005). Due in part to the various physical hazards brought by flooding, it is one of the deadliest types of natural hazards (Kellar and Schmidlin 2012). Hajat et al. (2005) found that most flood-related fatalities worldwide are due to the increased risk of drowning that comes with rapid-rise floods. Floods are also associated with an increased risk of accidental death due to trauma—people are at higher risk of being hit by objects in flowing flood water, for example—and vehicular fatalities; this is especially true in the United States (Ahern et al. 2005). In a study of vehicle fatalities caused by flash floods in Texas from 1959 to 2009, Sharif et al. found an increasing trend in the number of annual fatalities in Texas, and they found that while motor vehicle-related flash flood fatalities are a national issue, these fatalities in Texas are

greater than that in any other state during the same time period (Sharif et al. 2012). While floods of long duration were found to be potentially more important in the context of burden of disease (Ding et al. 2013), short-duration floods are more significant for accidental injury and fatalities (Spitalar et al. 2014). This is due, in part, to the resulting decreased time for warnings.

Other respiratory and cardiovascular-related health outcomes are of concern as well, during and following flood events. The dispersion of bioaerosols caused by favorable flood and post-flood conditions has led to concern about increased rates of respiratory illness (Solomon et al. 2006; Fisman et al. 2005). Pulmonary health can also be impacted by the floodwater itself—for example, direct injuries to the lungs can occur due to inhalation of water or traumatic injury (Robinson et al. 2011). Also, due in part to conditions like overcrowding and decreased access to quality health care, acute respiratory infections have been associated with natural disaster, including floods, with risks typically 3 to 5 days following the events (Robinson et al. 2011).

Solomon et al. (2006) investigated the effect of Hurricane Katrina on indoor and outdoor mold concentrations in New Orleans, Louisiana, and found that concentrations were considerably elevated 6 and 10 weeks after the hurricane. Increased exposure to mold is associated with increased respiratory irritation, as well as allergic or asthmatic responses—this is especially true for more susceptible people, including the elderly (Solomon et al. 2006). Additionally, while not statistically significant, Vanasse et al. (2016) found evidence that flooding may be associated with increased occurrence of cardiovascular disease, potentially due to the intense stress and unusual efforts brought about by flood events. The health effects of natural disasters can be particularly borne by emergency services workers—Tak and coauthors investigated the association between exposure to floodwater following Hurricane Katrina and

various health symptoms among firefighters. Of the 525 firefighters interviewed, 38% reported at least one new-onset upper respiratory symptom (Tak et al. 2007).

The environment in which a flood occurs can greatly mediate its impacts on health. For example, urban environments are more vulnerable to flash flooding, since various aspects of urban environments increase the volume and speed of flood runoff (Spitalar et al. 2014). Short and long-term effects are also impacted by characteristics such as the quality of infrastructure in an environment and the socio-economic status of those affected (Lowe et al. 2013).

The health effects of exposure to extreme precipitation and flooding can be both immediate and delayed. For example, many potentially dangerous pathogens found in contaminated drinking water or flood water have incubation periods of days or weeks (e.g., Colford 1999; Wade 2004; Falkinham et al. 2013; Farnham et al. 2014). Therefore, exposure to these pathogens may not result in noticeable health effects until a substantial amount of time after exposure. Similarly, exacerbations of existing, chronic health problems caused by exposure to extreme precipitation or flooding may not manifest in a deterioration in health on the same day as exposure, but days, weeks, or months later (Robinson et al. 2011; Vanasse et al. 2016). In the context of flooding, there are sustained health concerns related not only to the event itself but also to the restoration process, which can last long after the flood period (Ahern et al. 2005; Hajat et al. 2005).

Assessing exposure to extreme precipitation and flooding for epidemiological research

One key challenge in improving our understanding of the community-wide health risks associated with extreme precipitation and flooding is to better understand and improve exposure

assessment to these hazards within epidemiological studies. Previous studies have used a number approaches to assign exposure to extreme precipitation or flooding in epidemiological research.

Several studies have used community-specific measurements of precipitation and dichotomized days into those of extreme precipitation versus all other days. Thomas et al. (2006) in a study of extreme precipitation and waterborne disease outbreaks in Canada, chose to dichotomize daily rainfall using the 93rd percentile as a threshold for extreme events. A Sacramento-based study of precipitation and sick leave use also dichotomized precipitation measurements to identify exposed days, but used a much less stringent threshold for exposure, dichotomizing study days based on having any versus no precipitation (Colford et al. 1999). In their study of the association between asthma and extreme heat and precipitation, Soneja et al. (2016) used National Climatic Data Center (NCDC) meteorological data. The authors calculated county- and day-specific 90^{th} percentile thresholds, using 30-year baselines and 31-day windows, to calculate a binary exposure to extreme precipitation (Soneja et al. 2016). For example, to determine if there was extreme rain on a particular day in a particular county, the authors took the 90^{th} percentile of the distribution created by compiling all daily precipitation values for the month surrounding that day from 1960 to 1989 (Soneja et al. 2016).

Another study assessed exposure to extreme precipitation by incorporating rainfall beyond a community, extending their exposure analysis to identify relevant events in a community's entire watershed (Curriero et al. 2001). In this study, but the locations of GI outbreaks and weather station locations were coded to correspond to the center of the corresponding watershed (Curriero et al. 2001). Watersheds represent geographic units that drain all of the streamflow or rainfall to a common outlet. More importantly in the context of waterborne disease outbreaks, watersheds represent boundaries of drinking water sources

(Curriero et al. 2001). By incorporating watersheds in their analysis, Curriero et al. (2001) ensured that total monthly precipitation readings for a particular geographic area were being associated with waterborne disease outbreaks in the corresponding geographic area most likely affected by that precipitation.

In some cases, the exposure data of interest is included with information about health outcomes of interest. For example, in the review of weather-related motor-vehicle fatalities, the NHTSA's FARS dataset included information about the environmental conditions at the time of each crash—Ashley et al. (2014) focused on conditions that could affect visibility, including precipitation.

Occasionally in the literature, studies investigating the health effects of extreme precipitation hypothesized pathways of increased risk that include increased streamflow or flooding (Colford et al. 1999; Tornevi et al. 2013; Thomas et al. 2006; Nichols et al. 2009)—using extreme precipitation as a surrogate for flooding is not necessarily appropriate (Ivancic and Shaw 2015). In a Sacramento-based study of rainfall and sick leave use, rainfall was meant to act as a surrogate measure of exposure to combined sewer and storm water outflows, which, due to a corresponding increase in the exposure to waterborne pathogens, was hypothesized as the exposure likely to increase rates of illness (Colford et al. 1999). In their investigation of the association between sporadic cases of gastroenteritis and precipitation, Tornevi et al. (2013) used rainfall as a surrogate measure of outflows from combined sewer and storm water systems in Gothenburg, Sweden over 1,494 days. Their hypothesized pathway of increased risk of disease involved a decreased quality of river water used for drinking water production (Tornevi et al. 2013). Thomas et al. (2006) hypothesized an association between extreme rainfall and spring thaw conditions, measured using daily rainfall and daily streamflow, respectively, and

waterborne disease outbreaks in Canada from 1975 to 2001. These chosen exposures were meant to capture exposure to waterborne pathogens in the drinking water supply (Thomas et al. 2006). Similarly, Nichols et al. (2009) investigated associations between cumulative and excessive rainfall and water related disease in England and Wales from 1910 to 1999. Nichols and coauthors' (2009) hypothesized mechanism involved groundwater contamination with polluted surface water. Wash out from storm drains and contamination from runoff were also listed as relevant processes (Nichols et al. 2009).

These studies exemplify a few exposures common in literature examining gastrointestinal-related health impacts of extreme precipitation: contaminated drinking water due to combined sewer system outflows or runoff, or human exposure to the combined sewer system outflows or runoff itself. Hypothesized health impacts due to contaminated drinking water are valid: there have been several studies that have found associations between drinking water turbidity (i.e., cloudiness, which is used as a proxy for water microbial contamination) and gastrointestinal illness (Schwartz, Levin, and Goldstein 2000; Mann et al. 2007; Gaffield et al. 2003). However, the supposition that precipitation is an appropriate event on the pathway that leads to contaminated drinking water is less valid. Hydrologic models have predicted increases in both higher flood peaks and higher runoff with increased urbanization (EPA 1997; Brun and Band 2000); neither event can be described effectively with precipitation alone (Ivancic and Shaw 2015; Rowe and Villarini 2013). Rainfall is similarly a poor measure of human contact with combined sewer outflows or runoff. Flooding in urban areas, which increases the risk of human contact with combined sewer outflows or runoff, can occur when urban drainage systems deliver runoff at a rate faster than streams are able to transport it (Fisher et al. 1988). In this case, flooding is the more relevant event that leads to contact with outflow or runoff compared to

precipitation—while precipitation can sometimes lead to flooding, this relationship is not consistent nor reliable (Ivancic and Shaw 2015; Rowe and Villarini 2013), and it is possible for flooding to occur without precipitation (Groisman, Knight, and Karl 2001). Overall, the use of precipitation as a proxy for relevant exposures in the studies reviewed here is not entirely inappropriate, but in most cases using streamflow or the occurrence of floods would more effectively capture exposures of interest.

Flooding ascertainment varies between studies. Many studies examining the health effects of floods focus on a single extreme flooding event (Wade et al. 2004; Solomon et al. 2006; Tak et al. 2007; Setzer and Domino 2004; Ding et al. 2013). For studies investigating effects over time periods spanning multiple events, most either estimate flooding exposure using streamflow data (Thomas et al. 2006; Rowe and Villarini 2013; Garambois et al. 2015) or use datasets with human-entered flood events, such as the NOAA Storm Events database (Wade et al. 2014; Lin et al. 2015). There are similar datasets available for worldwide flooding events. For example, Hajat et al. (2005) obtained flood information from the Emergency Events Database (EM-DAT): The OFDA/CRED International Disaster Database, as did Jonkman et al. (2004). The EM-DAT database is publically available at www.emdat.be. In their review of the global health impacts of flooding, Ahern et al. (2004) also obtained flood information from this database. Events included in the database resulted in 10 or more deaths, 100 or more people reported affected, a call for international assistance, the declaration of a state of emergency, or any combination thereof (Hajat et al. 2005).

There are several limitations in estimating health effects related to extreme rain and flood events. Many studies to date have investigated the health effects of a single event, or have focused on a particular state or community, making it difficult to generalize to a larger

population (Jonkman et al. 2009; Tak et al. 2007; Fisman et al. 2005; Tornevi, Axelsson, and Forsberg 2013; Lin, Wade, and Hilborn 2015; Sharif et al. 2012; Setzer and Domino 2004; Wade et al. 2004; Wade et al. 2014). The use of different metrics to assign exposure may limit comparability between studies, especially if exposure metrics are poorly correlated.

Here, we evaluate and assess differences between four metrics for classifying community-level exposure to extreme rain and flooding. Additionally, using a time series analysis, we investigate the association between extreme precipitation and flood events and risk from four causes of mortality (accidental, respiratory-related, and cardiovascular disease-related) in 108 U.S. communities from 1987 through 2005. To date, the literature lacks a U.S.-based study of comparable scope investigating the mortality risks associated with extreme rainfall and floods. We hope to add to the literature by filling this gap, and by presenting a comparison of exposure metrics that could serve future studies investigating other health effects related to these extreme events.

CHAPTER 2

METHODS

Data

Study communities and time period

We conducted this study using 108 U.S. communities for which we had daily counts of mortality among residents from accidental, respiratory, and cardiovascular causes between 1987 and 2005 (Figure 1). Each community covered one or more U.S. counties, with a total of 124 counties making up the 108 study communities. The NMMAPS database was originally compiled to assess the health effects of five major ambient air pollutants across 20 of the largest U.S. cities from 1987 through 1994 (Samet et al. 2000). The database has since expanded to include data for 108 U.S. communities spanning from 1987 to 2005. This dataset includes daily exposure data for temperature, humidity, and several air pollutants and has been used extensively to study the human mortality risks associated in U.S. communities with exposure to several outcomes (Barnett, Huang, and Turner 2012), including mixtures of air pollutants (Roberts and Martin, 2006) tropospheric ozone (Bell et al. 2004), particulate matter (Samet et al. 2000), and temperature extremes (Anderson and Bell 2009; Anderson et al. 2013). This dataset has not previously been used, to our knowledge, to study the mortality risks associated with extreme precipitation or floods, nor has a study in the United States of similar scope been conducted with a different dataset. In this study, we generated daily exposure classifications for extreme precipitation and flooding for the 108 study communities in this database and joined our exposure metrics with the daily health outcomes by location and date to create a dataset to estimate acute associations between these weather phenomena and human mortality risk.

Precipitation data

We used two sources of precipitation data to identify extreme precipitation events in the study locations. First, we obtained daily precipitation data using the "countyweather" R package (Severson and Anderson 2016; Appendix A). This open source software package created as part of this Master's project and publicly available through the Comprehensive R Archive Network (CRAN). The "countyweather" package has been downloaded over 1,600 times by R users since it was published in October 2016. This package pulls data from the Global Historical Climatology Network (GHCN-Daily) of weather stations through the National Oceanic and Atmospheric Administration's (NOAA's) File Transfer Protocol (FTP) server. GHCN-D data is archived at NOAA's National Centers for Environmental Information (NCEI), and spans the 1800s to present. We obtained daily precipitation values (in millimeters per day) for each county in the study communities, and we then aggregated monitor-specific daily precipitation measurements across all stations in each study community to obtain a daily, community-level estimate of precipitation. This daily estimate was obtained using an unweighted average across all available stations on a particular day. During this aggregation step, we used only GHCN-D weather stations with non-missing data for at least 90% of the days in the study period: January 1, 1987 through December 31, 2005.

Second, we obtained daily precipitation data from the North America Land Data Assimilation System (NLDAS) Phase 2 through the Centers for Disease Control and Prevention (CDC) Wide-ranging OnLine Data for Epidemiologic Research (WONDER), an online health information system of the CDC (Mitchell et al. 2004). NLDAS Phase 2 is a collaborative project comprising precipitation, land-surface states, and fluxes from January 1979 to present. The CDC WONDER version of this NLDAS-2 precipitation data has been aggregated from the original,

gridded format of the NLDAS-2 data to county-level daily values for each U.S. county, and this county-aggregated version of the data is available from January 1979 through December 2011 (Mitchell 2004). From this data source, we obtained daily precipitation data (in millimeters per day) from January 1, 1987 through December 31, 2005 for all 124 U.S. counties within the study communities. In the case when more than one county comprised a community, a community-wide daily measurement was generated by averaging daily county-level precipitation measurements across all counties in the community.

Flood data

We also investigated two sources of data on flooding. First, we pulled streamflow data collected by the United States Geological Survey (USGS) from rivers and streams across the United States. To pull streamflow data for the study communities and study time period, we used the "countyfloods" R package, which pulls streamflow data using the United States Geological Survey (USGS) Water Services API while allowing users to query the data by date and county (Lammers and Anderson 2017). We pulled daily measures of stream discharge (cubic feet per second) from all stream gages with any available data in any of the counties belonging to the 108 study communities for January 1, 1987 through December 31, 2005.

Second, we pulled county-level flood data from the Storm Events database maintained by NOAA. The Storm Events database includes storms, weather events, and meteorological events which cause significant damage, loss of life, or injuries, are rare or unusual, or are otherwise significant. While most events in the database are based on reports by the National Weather Service (NWS), some are provided by outside sources (e.g., the media, individuals, or other government agencies) (Murphy 2016). While reports for some events in this database go back to

the 1950s, NOAA only began recording flood events in this database in 1996. We downloaded data for all years from 1996 through the end of our study period (2005) and pulled all events that occurred in a county in one of our study communities, limiting to events categorized with the keywords "Flood", "Flash flood", or "Coastal flood". We used the R package “noaastormevents” (development version available on GitHub) to facilitate downloading the data (Anderson and Chen 2017). Within the database, each flood event is recorded at the county level and with a begin date and end date—we used this information to generate a daily time series for each study community from January 1, 1996 through December 31, 2005 with a binary variable indicating whether a day was part of a flood event listed in NOAA Storm Events. In cases where more than one county comprised a community, we aggregated this flood data from county- level to community-level by assigning a day as a flood day for a community if at least one county in that community had a flood event recorded in the database on that day.

Mortality data

We obtained daily counts of accidental, respiratory-related, cardiovascular disease-related, and all-cause deaths across 108 U.S. communities from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) database (Figure 1). Daily mortality counts for each of the study communities were originally obtained from the National Center for Health Statistics. Deaths were classified by cause according to the ninth revision of the International Classification of Diseases (ICD-9)—mortality classifications included here include accidental (ICD-9 800–999), respiratory-related (ICD-9 490–496), cardiovascular disease-related (ICD-9 390–429), and all-cause (ICD-9 001–E999) (Samet et al. 2000). Here, we use an updated version of the NMMAPS dataset to evaluate associations between extreme precipitation, measured using NOAA and NLDAS data, and flood events, measured using USGS and NOAA Storm Events

data, and the four health outcomes listed above: accidental, respiratory-related, and cardiovascular disease-related deaths.

Daily temperature data (°F) for the 108 NMMAPS communities, which was included as a potentially confounding variable in the association between flood or extreme rain events and mortality outcomes, was also obtained from the NMMAPS database.

Classifying community exposure to extreme precipitation and flood days

Many epidemiological studies of the association between daily weather conditions and human health outcomes dichotomize exposure, especially in cases where the association between exposure and health outcomes might have a highly non-linear shape and, therefore, might be poorly modeled by a model assuming linearity between exposure and a function of risk of the health outcome. For example, Anderson and Bell (2009) investigated the health risks of heat waves as determined using 98th, 99th, and 99.5th percentiles of daily temperature distributions. Similarly, Liu et al. (2016) used 98th and 99.5th percentiles of particulate matter of 2.5 micrometers or less in diameter ($PM_{2.5}$) to determine the occurrence of "smoke waves" caused by wildfires. Several epidemiological studies of risks associated with extreme precipitation have similarly dichotomized daily precipitation values to identify extreme precipitation days (Thomas et al. 2006; Colford et al. 1999; Curriero et al. 2001; Groisman, Knight, and Karl 2001; Soneja et al. 2016; Ivancic and Shaw 2015; Pielke R.A. and Downton 2000, Georgakagos et al. 2014). While research studies often differ by specific choices made in dichotomizing continuous precipitation (i.e., some studies might choose to dichotomize by “any” vs. no precipitation; others might choose to use percentile thresholds), this general method of exposure ascertainment is relatively common in the literature. Here, we created a binary classification of extreme

precipitation exposure in each community from the continuous daily precipitation measurements obtained from each of the two precipitation data sources and used these to investigate agreement in binary classifications of daily extreme precipitation exposure across study cities, as well as to investigate how these binary classifications agreed with community-level daily flood exposure classification.

Within each study community and with each of the two sources of precipitation data, we identified days of extreme precipitation based on a threshold of daily precipitation. To calculate this threshold for NOAA and NLDAS precipitation data distributions across the study period, we first identified the 99^{th} percentile of daily precipitation values across all study days for each study community and then averaged these community-specific 99^{th} percentile values of daily rainfall across all study communities to obtain a single threshold to use to identify extreme precipitation days in all communities. Based on this calculation, we identified threshold values of 32.6 mm per day for NOAA data and 31.0 mm per day for NLDAS data. Any day in a community's dataset with precipitation exceeding the threshold value for a specific precipitation data source was classified as an extreme precipitation day.

We next used each of the sources of flood data to separately classify study days as exposed or unexposed to flooding within each community. For the USGS streamflow data, each gage measures the streamflow at single point in one river or stream, and the typical streamflow for any of these gages will vary widely by gage. Therefore, to use this data to identify flood days at a gage, we need to compare a gage's daily streamflow measure to a gage-specific threshold defining unusually high flow for that gage. To calculate these gage-specific streamflow flood thresholds, the median annual flood value (Q2) was calculated for each gage with at least 20 years of USGS annual peak flow data. This method is commonly used in flood-frequency

analysis, and involves taking the median value of yearly maximums over at least 20 years for a particular gage (Figure 2, Rao and Hamed 2000). Daily flood status for each gage was determined based on whether the daily streamflow exceeded the gage-specific flood threshold value. These calculations done by the "countyfloods" R package—output from this package gave binary flood values for each stream gage for each day of data pulled (Lammers and Anderson 2017). After identifying flood days at each stream gage with available data within each study community, we aggregated these values to a community-level flood exposure classification using the rule that if at least one gage in a county exceeded the flood threshold value on a given day, a flood event was recorded for that county on that day.

The NOAA Storm Events database is inherently binary, with a listing of flood events by county and with a start date and end date for each flood event (Murphy 2011). We expanded this into a time series of daily binary flood classifications, with all days from the start date to the end date of a flood event classified as exposed to flooding and all remaining days in the study period for a community classified as unexposed. We accounted for the change of Dade county, Florida (FIPS code 12025) to Miami-Dade county, Florida (FIPS code 12086), effective July 22, 1997, in processing this data.

We investigated regional patterns in these exposure metrics. For each exposure metric, we mapped the average number of flood days per year in each study community over the study period. We used the average per year to allow comparisons across the two flood data sources, since these two data sources have different periods of available data (1987–2005 for the NOAA ground-based monitor data and 1996–2005 for the NOAA Storm Events database flood listings).

Measuring correlation and agreement in exposure assessment across data sources

We next evaluated the correlation between precipitation exposure based on the two precipitation data sources considered, the NOAA ground-based stations and the NLDAS-2 Reanalysis data. First, we measured the strength of association between continuous daily precipitation measurements from the two data sources within each study community. Since the distribution of daily precipitation measurements within a community tends to be highly skewed rather than normally distributed, to measure this correlation we used two metrics of non-parametric rank correlation: Kendall's tau and Spearman's rho. Unlike Pearson's correlation coefficient, these rank correlation metrics do not require an assumption that either of the variables is normally distributed. Coefficient values closer to 1 indicate greater correlation between the two variables being evaluated. We calculated Kendall's tau and Spearman's rho coefficients within each of the study communities and then created summaries of these community-specific measurements across all study communities. Kendall's tau and Spearman's rho are two of the most commonly used nonparametric measures of association used (Fredricks and Nelsen 2007). These statistical values were chosen because they are nonparametric: they do not require data to conform to the strict assumptions required by Pearson's *r*, a more popular correlation statistic, such as a bivariate normal distribution (Chen and Popovich 2002). The two rank correlation coefficients are similar, often leading to the same statistical inference, but not identical—both range from -1 to 1, but should not be compared to each other (Gilpin 1993). Spearman's rho measures the extent of a monotonic relationship between two variables through the assignment of relative ranks to each pair of data (Spearman 1906), while Kendall's tau measures the number of pairs in two sets of ranked data that are in different orders (Kendall 1938; Sanderson and Soboroff 2007). The two tests differ in how they handle ties in the data as

well; Kendall's tau is generally better-suited to handing ties compared to Spearman's rho (Gilpin 1993). Because continuous precipitation data is expected to have a large number of ties in the ranked data due to large numbers of zeros, differences in how each test handles ties could result in important differences in calculated values of Spearman's rho and Kendall's tau. Therefore, we calculated both Spearman's rho and Kendall's tau for each test of correlation between sets of continuous daily precipitation data.

We used scatterplots to visualize patterns in continuous daily measurements of precipitation from the two sources of precipitation data for the six largest-population study communities, as well as for any communities with outlier values in the Kendall's tau and/or Spearman's rho distributions. In order to investigate differences in correlation in continuous precipitation metrics across communities, we calculated the extent to which Spearman's rho was correlated with (1) the number of NOAA weather stations contributing to average precipitation values, (2) the number of NLDAS observations (i.e., grid points included from the original, gridded NDLAS product when aggregating to county level), and (3) population in each community.

Next, we compared agreement in binary classifications of daily exposure to extreme precipitation within a community for the two precipitation data sources for exposures identified using nation-wide 99th percentile thresholds of precipitation. While total agreement in classification, as measured by the percent of days where classification agrees over the total number of study days, may seem to be an intuitive measure of agreement in exposure classification, it may be unhelpful in assessing agreement in exposure classification when exposure is rare. In the case of rare exposures, most days will be classified as unexposed by both metrics, and so this basic measurement of agreement can be very close to 1 (perfect agreement)

even in cases where the days classified as exposed to the hazard are completely different for the two exposure metrics. To illustrate this point, Figure 3 shows the two-dimensional distribution of daily measures of the two sources of precipitation data (values from NOAA ground-based stations are on the x-axis vs. values from NLDAS-2 Reanalysis data on the y-axis) for one of the study communities, Los Angeles, California. In each region of the plot, the color of the hexagon shows the count of days over the study period with those precipitation values. The vertical and horizontal black lines show the relative thresholds used to identify extreme precipitation in Los Angeles for each data source, and the letters identify four quadrants of exposure classifications: days classified as exposed to extreme precipitation by both metrics (quadrant A), days classified as exposed by NOAA ground-based monitor data but not by NLDAS-2 Reanalysis data (quadrant B), days classified as exposed by NLDAS-2 Reanalysis data but not by NOAA ground-based monitor data (quadrant C), and days classified as unexposed by both data sources (quadrant D). The vast majority of days in the study period were in quadrant D, indicating that the two precipitation data sources frequently agreed that there was not an extreme rain event on a given day. However, if the two precipitation data sources disagree on which rare days are exposed to extreme precipitation (i.e., if there were few or no days in quadrant A), epidemiological effect estimates based on binary exposure classifications could differ substantially depending on which data source was used for exposure classification, even if the overall agreement in exposure classification is strong because of the rarity of exposure.

Instead, we calculated agreement between binary measures of exposure to extreme precipitation by using the Jaccard coefficient (Jaccard 1912). The Jaccard coefficient excludes days classified as unexposed by both metrics when measuring agreement and so can provide a

more useful estimate of whether the identification of a rare exposure is similar across metrics. In relation to Figure 3, the Jaccard coefficient is calculated as:

$$J = \frac{a}{a + b + c}$$

where *a* is the number of days in quadrant A (where both metrics of interest detect an event), and *b* and *c* are the number of days in quadrants B and C, respectively (where one but not both of the metrics detect an event). The Jaccard similarity coefficient is particularly suited to evaluating the agreement between the exposure metrics we considered here because it does not incorporate the days in quadrant D of Figure 3, for which neither metric of interest detects an event. Extreme rainfall and flood events are relatively rare events—the Jaccard similarity coefficient prevents an inference of agreement due to many days that agree due to a lack of events. The Jaccard coefficient ranges from 0 to 1, with values closer to 1 suggesting higher similarity between the two metrics.

To investigate possible factors that might explain variation in Jaccard similarity coefficients across communities, we assessed correlation between Jaccard coefficients and (1) the number of stations inputting to the daily NOAA value on average, (2) the number of NLDAS observations, and (3) community population. We also visualized geographic patterns in Jaccard similarity coefficient values.

To compare agreement between community-level daily flood classification between the two sources of flood data (USGS streamflow data and NOAA Storm Events database), we similarly measured a Jaccard coefficient of similarity within each community, using the total number of days classified as a flood by either data source as the denominator and the number of

days classified as a flood by both data sources as the numerator; days for which neither data source identified a flood in the community (the majority of days) were excluded from the calculation. Since data on floods is only available since 1996 from the NOAA Storm Events database, we limited study data to the period 1996–2005 before calculating the Jaccard coefficient for agreement between the two sources of flood data considered for each community. To investigate factors that might explain differences in Jaccard similarity coefficient values for flood events across communities, we calculated the correlation between Jaccard values and (1) the number of USGS streamflow gages reporting in each community, and (2) community population. We also investigated geographic patterns in Jaccard values across communities.

Measuring the agreement between extreme precipitation and flood exposure

Next, we investigated whether extreme precipitation tended to occur on the same day or precede flood events by at most two weeks within each of the study communities. If measures of exposure to extreme precipitation and flooding are well-correlated within a community, then epidemiological studies could potentially use one exposure measurement as a surrogate for the other. Conversely, if the daily exposures to these two hazards are not well-correlated within a community, epidemiological studies need to be very rigorous in clarifying pathways by which a health outcome of interest might result from one or both of the exposures and be careful in selecting how to measure community exposure for time series studies.

To assess the correlation between exposure measurements, we first looked at a case study of documented flooding in two of our study cities. We examined the period of extreme precipitation and flooding resulting from an El Niño Southern Oscillation event that impacted two of the study communities in the winter of 1997-1998: Santa Ana / Anaheim, CA, and Los

Angeles, CA. Within each of these study communities, we investigated patterns in reported measurements of daily precipitation and flooding based on each of our data sources, to investigate whether exposure data were consistent during a major flooding event.

We next expanded to investigate agreement between flood exposure and extreme precipitation within all of our study communities across the full study period. To simplify this assessment, we used only one source of precipitation data (NLDAS) and one source of flooding data (USGS streamflow data). To compare the agreement between binary classifications of extreme precipitation and flood exposures, we calculated the following proportion *P* within each community:

$$P = \frac{r_l}{f}$$

where f is the number of days recorded as experiencing a flood in the community and r_l is the number of days in that set of community flood days when there was an extreme precipitation l days before the flood day. This metric measures the percent of days classified as exposed to flooding for which the community was also classified as exposed to extreme precipitation. We measured this value for lags of 0 to 14 days (l of 0 to 14) and also measured the proportion of flood days in a community for which an extreme precipitation day was measured for *any* of the days from lag 0 to 14. Lags up to two weeks were considered because heavy rainfall might lead to flooding several days after the fact.

To explore possible factors that could explain variation in proportion (*P*) values across communities, we investigated patterns in (1) average number of floods per year, (2) average length of floods (in days), and (3) geographic patterns.

Measuring the association between mortality risk and exposure to extreme precipitation and floods

Finally, we estimated the acute association between exposure to extreme precipitation and flood events and four mortality outcomes in the 108 study communities. To estimate this association, we fit a generalized linear distributed lag model to a daily time series of binary exposure (either extreme precipitation or flood) and daily mortality counts in each community, including control for day of the week, temperature, and long-term and seasonal trends (Gasparrini 2014; Zanobetti et al. 2000). The model included a distributed lag function, which allowed us to estimate the association between an exposure and a health outcome for several days following the exposure, while limiting potential problems introduced by collinearity in lagged exposures (Gasparrini 2014; Zanobetti et al. 2000). The distributed lag framework estimates a cumulative exposure over several days by constraining the parameter estimates for lags to follow a smooth pattern described by fewer coefficients than the number of lag days of exposure modeled. We included exposure lagged up to 14 days. This time period was chosen in order to capture both immediate and short to mid-term health effects of exposure to extreme precipitation or flooding. This modeling framework has been used in a large number of environmental epidemiology studies investigating associations between community-level health risks and ambient exposures, including air pollution and extreme temperature (Zanobetti et al. 2000; Xiao et al. 2017; Cox et al. 2016).

The equation describing this model we used is:

$$\log(E[Y_t^c]) = \beta_0 + \sum_{\ell=0}^{L} x_{t-\ell}^c \beta(\ell) + \sum_{j=1}^{6} \gamma_j dow_{jt} + s(T) + f(t)$$

where:

- Y_t^c is the count of mortalities on day *t* in community *c* for a certain cause of death (accidental, respiratory-related, CVD-related, or all-cause), where this count is assumed to follow a quasi-Poisson distribution, allowing for potential overdispersion in these daily counts;
- β_0 is the model intercept;
- $\sum_{\ell=0}^{L} x_{t-\ell}^c \beta(\ell)$ is the summation across all lag days of the indicator variable $x_{t-\ell}^c$ (whether there was a flood or extreme precipitation event at lag ℓ from day t) and the lag-specific coefficient (log relative risk at a specific lag for an exposed compared to unexposed day) is $\beta(\ell)$, which is defined using a distributed lag function constrained to follow a smooth pattern ($\beta(\ell)$ = *ns(ℓ, 3 df)*, i.e., the coefficient is determined based on a natural cubic spline function of the lag day, with three degrees of freedom in the spline, and with the function parameters identified through fitting the model to the observed data);
- $\sum_{j=1}^{6} \gamma_j dow_{jt}$ incorporates control for day of week as a factor, with separate model coefficients (γ_j) fit for each day of the week (dow_{jt}) with Mondays set as the reference weekday and so incorporated in the model intercept term;
- $s(T)$ is a smooth function of temperature, included to control for daily temperature, which we modeled using a distributed lag non-linear function; and
- $f(T)$ is a smooth function of time, included to adjust for long-term seasonal trends in expected mortality counts, which we modeled using a natural cubic spline with 7 degrees of freedom per year.

From this model, we calculated the cumulative log relative risk associated with a day of exposure over the period from the day of exposure to two weeks following the exposure. We conducted this analysis using the R package "dlnm" (Gasparrini 2011).

We fit this model separately within each study community to each pairwise combination of the exposure metrics of interest (specifically, binary exposure classifications determined using NOAA precipitation, NLDAS precipitation, USGS flood events, and NOAA flood events) and the mortality outcomes of interest (accidental, respiratory-related, cardiovascular-related, and all-cause mortality). Models run with NOAA flood events as the exposure of interest were fit to data only from 1996 though 2005 due to availability of NOAA flood data; all other models were fit to data from 1987 to 2005.

To estimate pooled overall effects of extreme rain and flood events on mortality, aggregated across all study communities, we combined community-specific estimates of the association between the rain or flood exposure and each morality outcome using a hierarchical model. We performed this aggregation using two-level normal independent sampling estimation (TLNise), a hierarchical model that assumes: (1) that community-specific estimates of log relative risk, estimated from fitting the model described above, are independent across communities, (2) that the estimated log relative risk for each community comes from a normally distribution centered on the true log relative risk in the community, and (3) that the true community-specific log relative risks follow a normal distribution centered on the overall, community-wide log relative risk (Peng and Dominici 2008). For this method of hierarchical pooling, communities for which parameter estimates were fit with greater confidence in the first-level, community-specific model exert a heavier influence on the overall estimate than communities for which community-specific parameter estimates were estimated with more uncertainty (Peng and Dominici 2008). We fit the

two-level normal independent sampling estimation using the "tlnise" R package (Everson and Morris, 2000). Since this package is no longer available on CRAN, we downloaded the archived source code and built the package from source locally.

CHAPTER 3

RESULTS

The populations of the 108 study communities, based on the 2000 Census, ranged from approximately 150,000 people to approximately 9.5 million (Figure 4). Across all study communities, average daily all-cause mortality counts ranged from around two deaths per day (Anchorage, AK) to around 180 deaths per day (New York, NY), with a median value of around 11 deaths per day (Tables 1 and 2). Among the specific causes of death considered, cardiovascular-related deaths were most common (median of approximately 5 cardiovascular deaths per day across the study communities) and least common for accidental deaths (median of less than 0.5 deaths per day) (Table 1).

Exposure characterization

NOAA and NLDAS precipitation

Precipitation data from NOAA ground-based stations was available for 105 of the 108 study communities for at least some part of the study period. The number of stations contributing to community-wide daily precipitation measures varied by community (Figure 5). For six communities, daily precipitation values were based on a single monitor, while for 14 communities, daily precipitation values were based on 10 or more stations on average per day. Los Angeles, CA had the highest average number of stations contributing to daily values (38.6).

For seven communities, some of study period did not have available data; these communities were still included in the study. For four of these communities, missing days were very sparse: Boston, MA, Washington, DC, Olympia, WA, and Kansas City, Kansas were

missing 2, 3, 3, and 27 days out of 6,940 total, respectively. Three communities were missing a higher percent of days: Jersey City, NJ, St. Louis, MO, and Baltimore, MD were missing 228, 491, and 1,130 days out of 6,940 total, respectively. Data were unavailable during the study period for three communities: Honolulu, HI, Richmond, VA, and Newport News, VA. For the other 98 study communities, precipitation data was available from NOAA weather stations for all days between January 1, 1987, and December 31, 2005.

Monitor locations were not evenly spaced across a community; instead, stations were often available at sites like airports, which could often result in stations being somewhat removed from the population center of the community. For example, Figure 6 maps the locations of available stations for Topeka, Kansas; of the five stations available for this community during the study period, the two stations labeled "A" and "B" are located at airports.

The NLDAS-2 Reanalysis precipitation data was, conversely, available for every study community in the contiguous United States over the entire study period (data was not available for Honolulu, HI or Anchorage, AK). Since the CDC WONDER database version of the data is aggregated at county level, there was a single daily measurement of precipitation available for each county in the study communities. This data was aggregated for the CDC WONDER database to county-level in a way meant to represent county-wide precipitation, rather than precipitation at specific points within the county (Mitchell et al. 2004). The daily number of values aggregated to generate a community-level daily measurement equaled the number of counties in the community, which ranged from one to six counties.

In most communities, correlation was moderate to strong between continuous daily precipitation measurements from the two sources of precipitation data. Figure 7 shows

scatterplots of daily precipitation measures from the two precipitations data sources in the six largest-population study communities: Chicago, Illinois, Dallas/Fort Worth, Texas, Houston, Texas, Los Angeles, California, New York, New York, and Phoenix, Arizona. Each point shows a day in the study, with the daily precipitation measure from NOAA ground-based stations on the x-axis and that from the NLDAS-2 Reanalysis dataset on the y-axis. If the two precipitation datasets were perfectly correlated, all points would fall along the diagonal reference line; in these six largest cities, the points generally fall close to this line, indicating moderate to strong correlation. However, while daily values show moderate to strong correlations between the two data sources, precipitation values from NOAA stations are consistently lower than NLDAS precipitation values in these six largest communities based on loess smoothing functions modeled to this data (smooth curves in Figure 7). The Kendall's tau coefficients of correlation between daily precipitation values from the two data sources range between 0.60 (Houston, TX) and 0.69 (New York, NY) for these communities, while the Spearman's ' coefficients range between 0.69 (Los Angeles, CA) and 0.80 (New York, NY). Both sources of daily precipitation data demonstrate that daily precipitation measures tend to be strongly right-skewed, with most days having no or little precipitation (lower left corner of the plots in Figure 7), and occasional daily precipitation values that are very high.

Expanding to all study communities, the measured rank correlations between daily measures of precipitation from the two precipitation data sources (NOAA stations and NLDAS-2 Reanalysis data) are fairly strong, with a median community-level Spearman’s rho of 0.75 and median community-level Kendall’s tau of 0.64 (Figure 8). In a few communities, the correlation in daily measures from the two sources is somewhat lower, with a Kendall’s tau below 0.5 in Kansas City, KS, St. Louis, MO, Johnstown, PA, and Washington, DC; conversely, correlation

was particularly strong in Portland, OR. Daily measurements of NOAA vs. NLDAS precipitation were compared for the six communities that were outliers in the Kendall's tau and or Spearman's rho correlation distributions (Figure 9).

We investigated several factors to determine if they help explain variation across the study communities in correlation between daily precipitation values from the two data sources. The correlation within a community between the two continuous daily precipitation measures was not correlated with the number of stations from which the NOAA monitor-based precipitation measurement was aggregated for the community, nor the number of NLDAS observations (i.e., grid points included from the original, gridded NLDAS product when aggregating to county level) contributing to the aggregated community value (Figure 10, left and middle panels). The correlation was not correlated with the population of the community (Figure 10, right panel).

To determine a threshold to use to identify extreme precipitation days, we measured the 99^{th} percentile of daily precipitation values in each community for each of the two precipitation data sources and then averaged these community-specific values to generate a single threshold to use for each data source. For NOAA precipitation data, the threshold we identified for extreme precipitation days was 32.6 millimeters; for NLDAS precipitation data, it was 31.0 millimeters.

The number of days of exposure to extreme precipitation varied substantially across communities, from an average of less than 0.05 days of extreme precipitation exposure per year (minimum average yearly exposure in Bakersfield, CA, and El Paso, TX, for the NOAA ground-based monitor data and in Spokane, WA, and Tucson, AZ, for the NLDAS-2 Reanalysis data; Figure 11, Table 2) to over 11 days of extreme precipitation exposure per year (maximum

average yearly exposure in Lake Charles, LA, for the NOAA ground-based monitor data and in Baton Rouge, LA, for the NLDAS-2 Reanalysis data; Figure 11, Table 2). Average exposure was typically highest in the Southeast and lowest in the Southwest. Spatial variation in number of days of exposure was smoother when exposure was assessed using the NLDAS- 2 Reanalysis dataset, for which it was very rare to have very dissimilar average number of exposed days in communities that were geographically close. By comparison, for exposure assessed based on the NOAA ground-based stations, there were a few cases where nearby cities had fairly different days of exposure per year; for example, Kansas City, KS falls into the third-highest sextile of number of exposed days while Kansas City, MO, falls into the highest. For exposure based on NLDAS-2 Reanalysis data, these cities both fall into the fourth-highest sextile of average number of days exposed to extreme precipitation (Figure 11).

Across the study communities, there was strong correlation between the two precipitation data sources in the number of days the community was exposed to extreme precipitation based on the two sources of precipitation data (Spearman's rho = 0.91, Table 1, Figure 12). Orlando, FL stands out as a relatively unusual outlier between average exposed days from NOAA and NLDAS data, with an average of 9.4 exposed days per year according to NOAA data and 5 exposed days per year according to NLDAS data. Orlando's relatively low number of NOAA stations contributing to its daily precipitation measurements (1.5 stations) could explain some of this lack of correlation.

While there tended to be strong correlation between the two precipitation sources in continuous daily precipitation metrics (Figure 8) and number of days a community was exposed to extreme precipitation (Figure 12), agreement was slightly dampened when considering agreement in the specific days identified as exposed to extreme precipitation within a

community. When we measured the Jaccard similarity coefficient to compare agreement in the days classified as exposed to extreme precipitation, we found the value was less than 0.50 in most study communities (Figure 13, upper left panel). This result indicates that, in most study communities, less than half of the days classified as exposed to extreme precipitation based on data from one of the precipitation data sources was classified as exposed based on both sources. New York, NY, had the highest agreement ($J = 0.68$). Conversely, in some communities, there was no agreement in the days identified as extreme precipitation days between the two precipitation data sources ($J = 0.00$; Bakersfield, CA, El Paso, TX, Phoenix, AZ, San Bernardino, CA, and Spokane, WA). In these communities, in other words, the two precipitation data sources identified completely non-overlapping sets of days as exposed to extreme precipitation.

We investigated a few possible factors that might help explain variation across communities in this Jaccard coefficient but found none of the factors were strongly related. Specifically, community-specific Jaccard coefficients of similarity in the days classified as exposed to extreme precipitation were not associated with (1) number of stations inputting to the daily NOAA value on average (Figure 13, lower left panel), (2) number of NLDAS observations (Figure 13, lower middle panel), (3) community population (Figure 13, lower right panel), or (4) geographic patterns (Figure 13, upper right panel).

USGS and NOAA flood events

USGS streamflow data was available for at least part of the study period for 92 out of 108 study communities from 1987 through 2005. For 81 of these communities, streamflow data was available for every day in the study period (1987–2005). For communities with streamflow data,

the number of gages contributing to the community-wide daily flood assessment varied (Figure 14), with a minimum of 1 average gage per day (St. Louis, MO, Newark, NY, Arlington, VA, Cedar Rapids, IA, Lafayette, LA, and Evansville, IN) to a maximum of 71 average gages per day (Seattle, WA). NOAA storm events information was available for all study communities from 1996 though 2005. Because of this disparity in study communities and years with available data, comparisons were made between the 92 communities with available USGS data from the years 1996 through 2005 with available NOAA flood data.

There was large variation in the average days of flood exposure, both across communities and between the two sources of flood data (Figure 15). Across communities, the distribution of exposed days per year is heavily skewed right. There were no strong regional patterns in average days of flood exposure based on either of the sources of flood data.

The distribution of number of days of flood exposure differed between the two flood data sources, with communities typically having a little more than twice as many flood days when exposure was assessed using USGS flood data compared to NOAA flood data (Table 1). There were more communities assessed to have very high exposure when exposure was classified using the USGS streamflow data compared to the NOAA storm events data—26 communities fall into the highest sextile of exposed days per year based on USGS data, while only 7 NOAA communities fall into this sextile (sextile cut-offs were determined separately for both exposure data sources). Average community-level exposure was poorly correlated for exposure classifications based on the two flood data sources considered—average yearly exposure based on USGS data was typically much higher in a community than average yearly exposure based on NOAA data (Spearman's rho = 0.07, Table 1, Figure 16).

The exact days classified as flood days within a community also disagreed substantially between the two sources of flood data, as measured Jaccard similarity coefficients for each community comparing flood classification based on USGS streamflow data and based on NOAA Storm Events data (Figure 17, upper left panel). The majority of communities had a Jaccard coefficient below 0.1, suggesting very poor agreement between the two measures of flood exposure, and all communities had Jaccard coefficients below 0.3, indicating there were no communities in which 30% or more of the days identified as a flood day by one flood data source were identified as flood-exposed by both data sources. In other words, on a day when the USGS streamflow data indicated a flood in a community, it was typically more likely than not that NOAA Storm Events database did not record a flood, and vice versa. This extreme lack of agreement in exposure classifications for flood exposure contrasts with our results for agreement across two data sources in extreme precipitation classifications, for which the Jaccard coefficient values suggested mild to moderate agreement in most communities.

The number of USGS gages contributing to average daily streamflow did not appear to be related to the community's Jaccard coefficient (Figure 17, lower left panel), nor did the community population (Figure 17, lower right panel) or geographic location (Figure 17, upper right panel). This lack of agreement between the two flood metric could in part be due to the average difference in length of flood. The average USGS flood lasts 5.5 days, while the average NOAA flood only lasts 1.3 days. Figure 18 illustrates a typical disparity between USGS and NOAA floods. This community had only one day recorded in the NOAA Storm Events database from 1996 through 2005, and a total of 202 days of exposure to USGS-derived floods during this period. Many of these USGS-derived floods lasted for several days or weeks (Figure 18).

Extreme precipitation and flood events

To compare measurements from our data sources on flood exposure and extreme precipitation exposure within a community, we first investigated as a case study a known period of extreme precipitation resulting from an El Niño Southern Oscillation event in 1997–1998 in two of the study communities as a case study, Santa Ana / Anaheim, CA, and Los Angeles, CA. These comparisons give greater insight into how extreme weather events are captured by the different exposure metrics. The El Niño Southern Oscillation event that occurred in 1997-1998 was one of the most powerful in recorded history—the event lead to several natural disasters worldwide, including in southern California (McPhaden 1999). We examined all four sources of exposure data (NOAA stations and NLDAS-2 Reanalysis data for precipitation, USGS streamflow data and NOAA Storm Events data for flooding) for Santa Ana/Anaheim, CA, from December 1, 1997 through February 1, 1998 (Figure 19), and for Los Angeles, CA, from January 15, 1998 through March 1, 1998 (Figure 20) in order to capture the time periods for which the effects of the El Niño were most extreme.

December 1997 in Orange County, California (the county comprising the Santa Ana/Anaheim community) was one of the wettest in history—the day of heavy rainfall on December 6, 1997 that resulted shows up in our data as a day when both NLDAS and NOAA data recorded an extreme precipitation day. The NOAA storm events database also includes December 6th as a flood day. Out of the eight USGS streamflow gages contributing data for Santa Ana / Anaheim during this period, seven of those did not record streamflow high enough to exceed the gage-specific thresholds (two of those gages, "11078000" and "11088500", are shown in Figure 19). Only one gage ("11047300") recorded a streamflow on December 6th that was high enough to exceed the gage-specific threshold of 1,820 cubic feet per second to result in

a USGS flood being recorded. There were also later, smaller peaks in streamflow later in December of 1997 and January of 1998; however, there was not a corresponding NOAA flood recorded on those days, nor enough rain reported by NOAA nor NLDAS to exceed the nation-wide 99th percentile thresholds we used to assess exposure to extreme precipitation.

The 1997-98 El Niño event affected Los Angeles, California in February of 1998 (Figure 20). In this period, there is more discrepancy between NOAA and NLDAS extreme rainfall, as well as between USGS and NOAA flood events. Similarly to Santa Ana/Anaheim, there were also eight USGS streamflow gages contributing to the USGS flood measure for Los Angeles during this period. Of these eight gages, five recorded streamflows high enough to exceed the gage-specific thresholds— again, a flood was recorded for a community on a particular day if at least one gage recorded streamflow that exceeded its threshold on that day. In early February, the NOAA Storm Events database recorded flood events; these events are reflected by either NOAA or NLDAS extreme rain days—there is one day in this period in early February when the two rain metrics agree. During this period, one stream gage had a high enough streamflow to classify as flooding ("11109395"), while the other two gages shown did not have correspondingly high streamflow. In late February, all three gages recorded high streamflow, and the NOAA Storm Events database also recorded two flood days. There was one day in late February when NOAA and NLDAS precipitation events did not agree, with NLDAS recording an extreme event when NOAA did not.

Next, we expanded our comparison of daily exposures to extreme precipitation and flooding to investigate all study communities over the full period. Specifically, we assessed whether flood- exposed days in a community tended to coincide with or closely follow days classified as having extreme precipitation. Across communities, there were generally very few

days with flooding that also had extreme precipitation on that same day—these exposures rarely coincide to occur on the same day (Figure 21, upper panel). There was some evidence that it is more likely for a flood day to coincide with extreme precipitation either on the same day or on one or more days in the previous two weeks (Figure 21, lower panel).

We investigated patters of proportions of flood days coinciding with precipitation with a two-week lag: Figure 22 shows results across our study communities of calculating the proportion of flood days (based on the USGS flood data) in a community for which the community also had an extreme rain event (based on the NOAA precipitation data) at lags from 0 to 14 days. This figure also includes columns for each community showing the average number of USGS flood days per community and the average length of floods identified in the community. The proportion of days with a flood event which also had extreme precipitation (*P*) was overall very low, close to 0% for most communities. A few communities show high proportions closer to 1 (i.e., close to 100% of flood days were associated with an extreme precipitation day in the community) for same-day comparisons (lag 0) or with a lag of one or two days. There did not appear to be a relationship between this estimated proportion *P* and the number of floods per year or the average length of flood across communities.

Communities were hierarchically clustered into four groups to determine if there were groups of communities with meaningful similarities in patterns of how often and at what lags flood days tended to be associated with extreme precipitation days. These clusters are displayed on the heatmap as row dendrograms (the colored branches shown on the left of the plot), with clusters of similar communities shown in neighboring rows of the histogram. The geographic locations of these community clusters is shown in Figure 23. There did not appear to be strong geographic patterns in the four clusters. For example, there were no geographical patterns to

suggest that the association between flooding and extreme precipitation exposure tended to be higher in areas of the country where flooding is likely to be rain-dominated rather than snowmelt-dominated.

Health impacts of extreme precipitation and flooding

Table 3 shows nationally-averaged risk estimates for the association between extreme precipitation (determined using NOAA and NLDAS rainfall data) or flooding (determined using USGS and NOAA Storm Events data) and accidental, respiratory-related, cardiovascular-related, and all-cause mortality. The effect estimates for all pooled effects indicated a slightly protective effect (relative risk below 1.00), with a few estimates barely achieving statistical significance. Overall, the pooled effect estimates suggest that, based on our exposure characterization and modeling choices, there is either a null or slightly protective overall association between risk of these causes of mortality and extreme precipitation or flooding.

Among community-specific estimates of the association between these two exposures and risks from these causes of death, there were a few statistically significant estimates but the vast majority of community-specific estimate were not statistically significant and were very close to a null association (relative risk of 1.00). Community-specific effect estimates for the association between NLDAS extreme rainfall and accidental fatalities illustrate this pattern (Figure 24). Further, effect estimates did not show geographic patterns, indicating that it is unlikely that overall pooling masked important effects within certain regions of the country (Figure 25). While Fresno, CA, and Boston, MA, had a statistically significant increased risk of accidental mortality associated with extreme precipitation, all other study communities show a statistically insignificant or protective effect, and we expect a few false positives for tests (i.e.,

erroneously rejecting the null hypothesis) given the number of communities for which we are modeling the association, so the few statistically significant results observed are likely spurious.

We also estimated associations between exposure to every other pairwise comparison of exposure (extreme precipitation and flooding) and the four health outcomes of all-cause, cardiovascular, respiratory, and accidental deaths. While there were a few communities for each measured association for which the association was statistically significant, some false positive results are expected given the number of communities for which associations were measured. We hypothesized a true positive association between extreme precipitation and flooding and the four health outcomes considered. This true association would have been evidenced by statistically significant pooled relative risk estimates greater than one. Our results do not support our hypothesized association; there is no evidence from the observed overall effect estimates to suggest that this true association is likely between either of the exposures and any of the health outcomes considered (Table 3). Furthermore, we found little evidence that there were regions of the U.S. in which estimated associations were consistently positive or negative. For each of the estimated associations, community-level estimates were poorly to-moderately correlated when extreme precipitation exposure was measured using the two sources of precipitation data considered (Figure 26, upper panels, Table 4). All Spearman's rho values for extreme precipitation-derived estimates indicated a positive correlation. Community-specific estimates were completely uncorrelated when USGS flooding exposure was measured against NOAA flooding (Figure 26, lower panels, Table 4). All Spearman's rho values for flood-derived estimates were negative and very close to zero.

CHAPTER 4

DISCUSSION

Overall, our results suggest reasonable correlation between the two precipitation data sources under consideration (NOAA monitor-based and NLDAS-2 Reanalysis data-based) for continuous daily precipitation measurements and estimates of annual days of exposure for a community, weaker agreement between the days classified as extreme precipitation days by these two precipitation data sources, and poor agreement between exposure classifications based on the two flooding measurements (USGS streamflow-based and NOAA Storm Events-based). In most communities, only a low percentage of flood days were concurrent with or preceded by 14 or fewer days by an extreme precipitation day. When we investigated the associations between extreme precipitation or floods and risks of all-cause, accidental, respiratory-related and cardiovascular-related mortality using any of the exposure data sources considered, we found these exposures tended to have no association or be associated with a slightly decreased risk of all outcomes. Estimates of this association were moderately correlated when comparing estimates made using the two precipitation data sources and uncorrelated when comparing estimates made using the two flood data sources. Our results provide insight for measuring exposure to extreme precipitation and floods in epidemiological research.

In past epidemiological and other research, various methods have been used to measure community-level exposure to precipitation and model its effects. Some studies have chosen to treat rainfall as a continuous variable (Drayna et al. 2010; Fisman et al. 2005; Thomas et al. 2006; Tornevi, Axelsson, and Forsberg 2013), while several other past epidemiological studies of precipitation and health risks have dichotomized continuous daily precipitation measurements

to identify days of extreme precipitation, as we do here (Thomas et al. 2006; Colford et al. 1999; Curriero et al. 2001; Groisman, Knight, and Karl 2001; Soneja et al. 2016; Ivancic and Shaw 2015; Pielke R.A. and Downton 2000, Georgakakos 2014). For these studies that dichotomize precipitation exposure, researchers have varied in how they define the binary extreme precipitation variable. Researchers have defined extreme precipitation days as events that occur once in a defined amount of years, or as days with precipitation exceeding a specific percentile of the distribution (Groisman, Knight, and Karl 2001). For studies that use percentile-based thresholds, different percentiles have been used. Soneja et al. (2016), for example, identified a day-of-year-specific 90^{th} and 95^{th} percentiles of precipitation for each county and for each day in their study period. An extreme precipitation day was recorded if the precipitation on a day exceeded its county- and day-of-year-specific threshold. Ivancic and Shaw (2015), in a study examining precipitation and river discharge trends, used a threshold of the 99^{th} percentile of days with precipitation across the historic period of record. This method, the authors note, is consistent with the Third National Climate Report (Georgakakos et al. 2014). Conversely, a few studies have used an absolute millimeter per day value to serve as a threshold**:** Karl et al. and Groisman, Knight, and Karl defined extreme precipitation as a day with precipitation above 50.8 mm (1996, 1999). Similarly, Pielke and Downton (2000) defined extreme precipitation as days with more than 2 inches (50.8 mm) of precipitation.

While most communities showed reasonable correlation between measures of daily continuous precipitation, correlation was low in a few communities (Figure 9). Particularly low correlation between the two sources of precipitation data was most notable in Kansas City, KS (tau = 0.43; rho = 0.50). A possible explanation for this discrepancy is the low number of NOAA stations contributing to daily values—Kansas City had two stations contributing to daily values

during the study period; however, one of those stations had missing data the majority of the time resulting in an average of 1.03 stations for the study period. It is possible that the values from this single monitor were biased based on geographic location compared to values from NLDAS-2 Reanalysis data. However, St. Louis, MO and Washington, DC, which had higher correlation between daily precipitation values compared to Kansas City, KS, also had a low average number of stations contributing to daily values (0.96 and 1.90, respectively). Across all communities there was no evidence of a link between number of stations and correlation between precipitation metrics (Figure 10, left panel). Therefore, it is not clear why the low correlation in Kansas City, KS is so striking compared to other communities with few NOAA stations.

Many of the epidemiological studies that use a threshold to define extreme precipitation base that threshold on the single city or state of their study population (Soneja et al. 2016; Colford et al. 1999; Tornevi et al. 2013). Here, we used a nation-wide threshold for all study communities to identify extreme precipitation. Future research could explore whether defining extreme precipitation days using relative thresholds, determined separately for each community based on its climate, modifies the associations observed here between (1) extreme precipitation and flooding and (2) mortality risks. Research about the health risks associated with extreme temperatures have found that community-specific thresholds for extreme event definitions are effective in identifying events to which a community may be poorly adapted (Buguet 2007; Nixdorf-Miller, Hunsaker, and Hunsaker 2006; B. G. Anderson and Bell 2009; Yang et al. 2017, USCGRP 2016); it is possible that a similar approach might identify stronger associations between extreme precipitation and mortality risk than those found here. Although many epidemiological studies of precipitation used cut points to determine exposure to extreme rain, our research suggests that using a continuous measure of precipitation may be more robust to the

precipitation data used. In environmental epidemiology health studies for which dichotomization is used, our analysis found only low correlation in effect estimates calculated using one versus the other of the precipitation data sources, suggesting that epidemiological results could be highly sensitive to the choice of precipitation dataset.

For flood exposure assessment, many previous studies have used one of the two methods investigated here—some researchers have relied on existing databases of defined flood events such as the NOAA Storm Events database (e.g., Wade et al. 2014; Wade et al. 2004; Lin, Wade, and Hilborn 2015; Kellar and Schmidlin 2012), others have used streamflow to measure dichotomized flood events. Thomas et al. (2006) modeled streamflow using a few different methods; one method was to use a rolling five-day cumulative average streamflow for each station—the maximum cumulative average six weeks prior to a waterborne disease outbreak was selected for analysis. Slater and Villarini (2016) measured flooding using USGS gage height values, which indicate water surface elevation. Four corresponding National Weather Service (NWS) numeric thresholds indicated categories of severity—action, minor, moderate, and major (Slater and Villarini 2016). Stephens et al. (2015) measured the extend of "floodiness" by calculating the percentage of gridded river cells that exceed a defined threshold in a given time period. Based on our results, epidemiological studies of the health risks associated with floods could be highly sensitive to the source of data used to assess flood exposure; we found almost no correlation between community-level associations between flood and mortality risk when using one versus the other of these flood data sources.

The discrepancy between the two flood data sources in the average number of exposed days per community may be due to the nature of flood determination—NOAA flood events are often reported based on the human impacts of a particular event, while USGS floods were

determined based on systematic streamflow threshold calculations. The former measure may be more selective, missing floods that did not have obvious impacts on humans or caused little damage. Additionally, the qualitative data comprising the NOAA Storm Events database could be biased in that submitted events included may be more likely to be located in more populated areas. Despite efforts of standardization, non-standard event types (i.e., event types not present in NOAA Storm Data documentation (Murphy 2011) are present in the database (dos Santos 2016). While these non-standard event types amount to only about one-fifth of a percent of the total number of events each year, non-standard labeling could result in some relevant events being missed in database searchers (dos Santos 2016). Relatedly, although each element of an extreme event is meant to be given a separate entry in the database, some floods during tropical storms may have been erroneously included within an event listing categorized as "tropical storms," and therefore would not show up in our NOAA Storm Events flood dataset. Furthermore, the USGS method tends to result in more long-lasting flood events compared to NOAA floods, so some of the discrepancy between the two data sources may also result from days later in a flood that are included based on USGS data but not for NOAA data. Figure 27 illustrates this phenomenon in Tampa, Florida from 1996 though 2005. While there are periods when both NOAA and USGS flood data sources detect a flood, it is common in this dataset for USGS floods to detect more days of exposure compared with NOAA floods, and for the USGS floods to last much longer than NOAA floods.

Some of the disagreement between the two sources of precipitation or flood data may additionally result from extreme precipitation or flood events being very localized. This makes it hard to adequately characterize exposure across a community, and could result in disagreement between two datasets if they are capturing exposure at different locations within the community.

For example, the locations of NOAA precipitation stations will typically not be identical to the locations of grid points used for NLDAS data. For many environmental epidemiology studies of community-level risks associated with ambient exposures, epidemiologists have used monitor-based exposure measurements to estimate community exposure, including for studies of risks associated with temperature and air pollution (e.g., Samet et al. 2000; Brook et al. 2010; Anderson and Bell 2009; Anderson and Bell 2011). In these cases, values measured at one or a few stations are often used as a metric of exposure for an entire community.

In contrast with some other environmental exposures such as temperature, which can be relatively spatially homogenous across a community, precipitation can be highly localized, making it difficult to characterize exposure throughout a community using observations from stations (Borga et al. 2014). In some locations extreme precipitation may be more likely to result from very localized events like thunderstorms, potentially affecting only a small part of the community, rather than regional storms that would affect all parts of the community. While monitor-based measurements might be appropriate for exposures that are spatially homogenous, for which exposure throughout the community is more likely to be similar to values measured at the monitor(s), they may be more problematic for assessing daily community exposure to precipitation for epidemiological studies. Although some epidemiological studies have identified health risks associated with extreme precipitation and flooding (Lin et al. 2015; Thomas et al. 2006), studies with finer spatial resolution for both health data and exposure data might more clearly identify relevant risks given this potential for spatial heterogeneity across a community in exposure.

Given this spatial heterogeneity in precipitation within a community, the NLDAS-2 precipitation data investigated here may provide improved exposure assessment for

epidemiological studies compared to monitor-based measurements. While NLDAS-2 Reanalysis data similarly provides community-level estimates, the method of producing these estimates differs from monitor-based data—NLDAS-2 data is produced from a combination of modeled and observed data (Mitchell 2004) and, while also only generated for points within the community, is aggregated at equally spaced grid points. Land surface models (LSM) have been continually improved upon since their introduction in the 1960s, helping researchers understand land surface-atmospheric interactions (Zhao and Li 2015). Data assimilation (i.e., the combination of LSM simulations with ground and satellite-based observations) can minimize the effects of errors in these models (Zhao and Li 2015). For example, the Global Land Data Assimilation System (GLDAS) incorporates multiple LSMs and a large quantity of observed data—this model produces the hourly 1/8 degree (14 by 14-kilometer square) geographic-area gridded data that comprises the North American Land Data Assimilation System (NLDAS) (Zhao and Li 2015). The number of observations per day that contribute to community-level NLDAS Phase 2 estimates in this study represent the total number of precipitation measurements (mm) recorded for the 1/8-degree geographic area grids comprising a study community.

Community-level NLDAS-2 precipitation data and monitor-based precipitation data both represent aggregations of smaller-scale data; however, differences in the inputs (i.e., NLDAS-2 data is a combination of modeled and observed data; NOAA-monitor-based data is comprised solely of observed data) could result in important differences in summary measures. Differences in summary measures between NLDAS-2 and NOAA monitor-based data could also arise due to the tendency of monitor-based data to have both sporadic and lengthy periods of missing values (Wilby et al. 2017). Furthermore, the data derived from stations can be inconsistent because of changes in the location of the site or of the monitor equipment (Wilby et al. 2017).

In the case of the NOAA monitor-based data, the representativeness of specific monitor locations in describing population exposure is also possibly problematic, given that the stations are often located away from the population center of the community (e.g., at airports). For epidemiological studies assessing the community-wide health risks associated with precipitation exposure, therefore, reanalysis data like the NLDAS-2 data available might provide a better estimate of exposure; although it also creates a community-wide estimate based on values at specific points in the community, its use of equally spaced grid points improves the epidemiological relevance of its summary measures. This data source has the added advantage of being available for all counties in the contiguous U.S. and for all days (currently between 1979 and 2011) (Mitchell et al. 2004).

Like precipitation, flooding can also be very localized, affecting only specific areas in a community. Overall exposure across a community is not consistently captured by stream gages because of the variation of stream gage density between and within states (Rowe and Villarini 2013). It is possible that scaling up flood events to the scale of communities (one or multiple U.S. counties) could prevent us from detecting an association between flooding and mortality, as many of the flood days identified by our methods may only have affected parts of their communities. We found that two sources of data relevant to flooding and used in previous epidemiological studies (Timothy J. Wade et al. 2014; Timothy J. Wade et al. 2004; Lin, Wade, and Hilborn 2015; Kellar and Schmidlin 2012; Ashley et al. 2015; Jonkman et al. 2009; Groisman, Knight, and Karl 2001) disagreed substantially in classifying days as exposed and unexposed within a community over the same time period and even disagreed strongly in the number of days a community was exposed to flooding on average each year.

Another critical result of our analysis for planning and interpreting epidemiological studies is the finding that neither metric of flooding was well correlated with exposure to extreme precipitation. While long-term, systematic increases in precipitation are predicted to cause general increases in streamflow, this relationship on a smaller temporal scale depends on other factors like soil moisture and snow cover (Groisman, Knight, and Karl 2001). Others have found this holds true for predecessor rain events (i.e., systems producing more than 100 mm of rainfall per day) and flooding—Rowe and Villarini (2013) found that these events cause extensive flooding in the Midwest, but also that rainfall is not a good proxy for discharge, because of discharge's dependence on soil moisture conditions and land use or land cover. Ivancic and Shaw (2015) similarly failed to find a consistent relationship between heavy precipitation and flooding, reporting that 99th percentile precipitation results in 99th percentile river discharge 36% of the time. Furthermore, there are conditions when flooding can occur without being preceded by heavy rainfall, including dam or levee failure and debris and landslide floods (Perry 2000). Groisman and coauthors investigated the link between precipitation and flooding by selecting precipitation and streamflow stations within the same 1-by-1 degree grid boxes; many other hydrologic studies investigate this relationship at the watershed level (Groisman, Knight, and Karl 2001, Ivancic and Shaw 2015). The lack of agreement observed here between extreme precipitation and flooding may be partially explained by the coarseness of community units in this study—incorporation of watershed units or comparison of precipitation and stream gages within the same grid boxes (instead of within county units, which can span several grid boxes) might somewhat improve observed correlations.

Despite the lack of a consistent relationship between heavy precipitation and increased stream discharge, Ivancic and Shaw (2015) observed that non-hydrologist researchers often

mistakenly use heavy precipitation as a surrogate for trends in flooding. This is possibly the case in a few epidemiological studies reviewed here (Colford et al. 1999; Tornevi, Axelsson, and Forsberg 2013; Thomas et al. 2006; Nichols et al. 2009) which were based on the hypothesis that precipitation was a relevant cause for contaminated drinking water due to combined sewer system outflows or runoff, or for human contact with combined sewer outflows or runoff. In the case when precipitation is hypothesized to lead directly to outflows, this hypothesis is not incorrect; however, when flooding is an intermediate step in the hypothesized mechanism (i.e., heavy precipitation leads to flooding, which leads to outflows or runoff), the use of precipitation is inappropriate. The use of precipitation data instead of streamflow or flooding data could be partially due to a misunderstanding of the relationship between precipitation and flooding among non-hydrologists, and partially due to the relative ease of obtaining precipitation data compared to flooding data—there are about half as many stream discharge measurements in the U.S. compared to precipitation measurements (Ivancic and Shaw 2015).

In future epidemiological studies it may be possible to improve estimates of associations between health risks and extreme precipitation or floods through a more localized, sub-community exposure characterization. Other data products could be explored for such studies—for example, the Federal Emergency Management Agency (FEMA) of the U.S. Department of Homeland Security offers several insurance-related data products related to flood risk built on the National Flood Insurance Program (NFIP). The NFIP was established in 1968 with the Housing and Urban Development Act, and offers flood insurance to communities that agree to implement flood mitigation measures (Vaughan 1997; Browne and Hoyt 1999). Insurance rates under the NFIP are based on residential proximity to Special Flood Hazard Areas (SFHAs) (Browne and Hoyt, 2000). Flood Insurance Rate Maps (FIRMS), which show SFHAs in a

community, are offered by FEMA as a flood insurance-related data product. Residential proximity to SFHAs, paired with data documenting dates of floods in a community like the NOAA flood events listing, could represent a spatially-improved exposure metric for flooding compared to USGS stream gage data or NOAA Storm Events data. This method of exposure assessment could be further explored for epidemiological research in a future study.

An additional measure of precipitation that future research could evaluate is the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR) dataset. Similarly to NLDAS-2 Reanalysis data, NARR data incorporates both land surface model output and precipitation observations (Mesinger et al. 2006). Additionally, advances in methods to downscale gridded model predictions, such as NLDAS-2 output, could make localized estimates of extreme precipitation exposure more accessible (Zhang 2005). While these spatially refined flood and precipitation metrics might give improved estimates of exposure, refined location data would also need to be available for each health event—this is not always the case.

Another avenue for future work related to this project involves seasonal analysis. Soneja et al. (2016) observed an increased risk of hospitalization for asthma associated with summertime extreme precipitation events. It is possible that extreme precipitation and flooding are similarly associated with an increased risk of mortality during a particular season.

The disconnect we observed between exposure profiles resulting from different precipitation and flooding data sources emphasizes the importance in environmental epidemiology studies of carefully choosing an exposure metric that describes the hypothesized pathway of exposure. We have shown that differing precipitation and flood data generate

exposure classifications of extremes that are not well correlated within or between exposure type. For example, when comparing relative risk estimates between models using the two precipitation data sources, we found community-level estimates were only moderately correlated, and that relative risk estimates of models using two flooding data sources were very poorly correlated (Table 4, Figure 27). These results suggest that community-level estimates are very sensitive to the choice of data used in the analysis. Therefore, for example, if researchers are interested in health outcomes related to flooding, they would be ill-advised to use extreme precipitation as a surrogate for that exposure, as in most large U.S. communities a measure of extreme precipitation will capture few of the days the community is exposed to flooding. For example, in an investigation of the association between precipitation upstream of a drinking water facility and acute gastrointestinal illness, the hypothesized mechanism of increased risk involved increased exposure to waterborne pathogens—this might have been better characterized by increases in streamflow or incidence of flooding rather than extreme precipitation (Tornevi et al. 2013).

When we investigated associations between extreme precipitation or flooding and mortality outcomes, we found a null or protective association between extreme precipitation and flooding and four different categories of mortality. A possible explanation for this protective effect in the context of accidental fatalities could involve behavior changes that occur as a result of extreme weather conditions. Hassan and Abdel-Aty (2011) reported that there are several human factors (e.g., longer driving experience, number of driving citations) that influence a driver's compliance with variable speed limits (VSL) and recommendations conveyed through changeable message signs (CMS) in reduced visibility weather conditions. Extreme rain and

flooding events could influence behavior by decreasing the chance of reckless driving, or of driving at all, contributing to the observed protective effect for accidental mortality.

We also considered the explanation that extreme precipitation may be protective because of an association with a reduction in other dangerous exposures such as air pollution. Several studies have demonstrated an association between air pollution and mortality (e.g., Pope and Dockery 2006; Samet et al. 2000; Brunekreef and Holgate 2002), so if precipitation reduces this exposure, this pathway may contribute to the observed protective association for all-cause, cardiovascular, and respiratory mortality. However, there is little evidence in the literature that extreme precipitation might consistently result in decreased levels of ambient air pollution—instead, researchers have observed that increased air pollution could result in decreased precipitation (Givati and Rosenfeld 2004; Rosenfeld et al. 2007).

We found null associations for most mortality outcomes. Null associations were found for accidental, respiratory-related, and CVD-related mortality for all four exposure metrics, excluding respiratory-related mortality with NLDAS-2 data, and CVD-related mortality with NOAA flood data. Associations for the remaining mortality outcomes and exposure combinations were statistically protective, but very close to null (Table 3). As noted previously, these null associations might reflect exposure misclassification in assigning exposure measured at certain points in a community to the full community population. The direction and magnitude of the bias potentially caused by this measurement error would depend on how severely exposure is misclassified (Rothman et al. 2008). For example, if for some of the days identified as exposed to extreme precipitation or flooding only a small portion of the community's population was actually exposed, it is possible that the entire community is recorded as being exposed to events that most of the community is not exposed to. It is unlikely that this over-estimation of exposure

would be differential by the health outcome (i.e., would differ between those in our dataset who did and did not pass away during the study period). Therefore, non-differential over-estimation of exposure would bias community-specific effect estimates towards the null (Rothman et al. 2008). This again highlights the potential importance of more localized exposure assessment for epidemiological studies of extreme precipitation and floods, and that studies based on community-wide exposure assessments might miss important associations.

In comparing two precipitation datasets and two flood datasets commonly used in the literature, we found moderate correlation between NOAA and NLDAS-2 continuous precipitation measures, and weaker agreement between extreme precipitation events derived from each dataset. We found poor agreement between USGS streamflow-derived floods and NOAA Storm Events floods. Our finding that precipitation rarely precedes flood events within two weeks agrees with hydrologic literature and theory, but suggests that epidemiologic studies studying health impacts of precipitation that hypothesize exposure pathways that include flooding may not adequately capture exposures of interest. We found null or slightly protective associations between all four exposures and accidental, respiratory-related, cardiovascular disease-related, and all-cause mortality, suggesting that extreme precipitation and flooding may be localized to an extent that their health impacts cannot be captured on the community level. These associations were moderately correlated between precipitation measures, and uncorrelated between flood measurements. Future research could further investigate differences in exposure profiles resulting from additional precipitation or flooding data sources. These comparisons will be increasingly relevant with climate change, which will mainly affect societies around the world through weather and climate extremes (Trenberth, Fasullo, and Shepherd 2015). There is an increasing demand on scientists for updated assessments of the impacts of extreme events; these

assessments will require an understanding of how measurements of extreme events can vary across data sources (Trenberth, Fasullo, and Shepherd 2015). Our findings have important implications for future epidemiologic studies investigating health outcomes related to extreme precipitation or flooding—relevant pathways in hypothesized exposures should be carefully considered, and should more exclusively inform exposure data choices.

TABLES

Table 1. Mean, median, and 25th and 75th percentile values for the distribution of the number of exposed days per year across study communities for rain (based on a threshold of the average of community-specific values of the 99th percentile of daily, year-round precipitation) and flood, as well as mean, median, and 25th and 75th percentile values for the distribution of fatalities per day across study communities for accidental, respiratory-related, cardiovascular-related, and all-cause mortality. NOAA flood events are measured from 1996 through 2005; all other exposure metrics are measured from 1987 through 2005.

Measure	Mean	Median	25th percentile	75th percentile
Average number of exposed days in communities				
Extreme precipitation exposure days				
NOAA monitor data	4.2	4.2	2.0	6.0
NLDAS-2 Reanalysis data	4.2	4.6	2.0	6.0
Flood exposure days				
USGS streamflow data	8.3	3.8	1.3	9.5
NOAA Storm Events data	2.9	2.4	1.3	3.9
Average daily mortality counts in study communities				
All-cause deaths	18.3	11.3	6.2	19.8
Accidental deaths	0.7	0.4	0.2	0.8
Respiratory deaths	1.6	1.1	0.5	1.8
Cardiovascular deaths	7.8	4.7	2.6	8.1

Table 2. Communities with minimum and maximum values in the distribution of the number of exposed days per year across study communities for rain (based on a threshold of the average of community-specific values of the 99th percentile of daily, year-round precipitation) and flood, as well as mean, median, and 25th and 75th percentile values for the distribution of fatalities per day across study communities for accidental, respiratory-related, cardiovascular-related, and all-cause mortality. NOAA flood events are measured from 1996 through 2005; all other exposure metrics are measured from 1987 through 2005. The value listed in parentheses after each community is the average number of exposed days per year for the four exposure metrics, or the average number of fatalities per day for the four mortality outcome categories.

Measure	Community with minimum (value)	Community with maximum (value)
Average exposed days per year		
Extreme precipitation exposure days		
NOAA monitor data	Bakersfield, CA, El Paso, TX (0)	Lake Charles, LA (11.32)
NLDAS-2 Reanalysis data	Spokane, WA, Tucson, AZ (0.05)	Baton Rouge, LA (11.47)
Flood exposure days		
USGS streamflow data	Arlington, VA, Baltimore, MD, Cincinnati, OH, and Newark, NJ (0)	Orlando, FL (62.63)
NOAA Storm Events data	Seattle, WA (0.10)	Tucson, AZ (13.50)
Average daily mortality counts in study communities		
All-cause	Anchorage, AK (2.067)	New York, NY (182.80)
Accidental deaths	Arlington, VA (0.05)	New York, NY (5.47)
Respiratory deaths	Anchorage, AK (0.16)	Los Angeles, CA (14.12)
Cardiovascular deaths	Anchorage, AK (0.6)	New York, NY (88.32)

Table 3. Pooled effect estimates of the overall association across the 108 study communities between four different exposure metrics and four categories of mortality. Relative risk values represent the posterior means of pooled, nation-wide effects. Models were adjusted for temperature, day of the week, and long-term and seasonal trends.

Exposure	Mortality outcome	Relative risk	95% posterior interval
NOAA precipitation	Accidental	0.981	(0.952, 1.011)
	Respiratory-related	0.981	(0.962, 1.000)
	CVD-related	0.995	(0.998, 1.002)
	All-cause	0.992	(0.987, 0.998)
NLDAS rainfall	Accidental	0.969	(0.938, 1.000)
	Respiratory-related	0.980	(0.961, 0.999)
	CVD-related	0.994	(0.986, 1.003)
	All-cause	0.990	(0.984, 0.995)
USGS flood events	Accidental	0.982	(0.957, 1.008)
	Respiratory-related	0.995	(0.978, 1.012)
	CVD-related	0.997	(0.989, 1.004)
	All-cause	0.995	(0.990, 0.999)
NOAA flood events	Accidental	0.977	(0.563, 1.694)
	Respiratory-related	0.996	(0.978, 1.014)
	CVD-related	0.990	(0.980, 0.999)
	All-cause	0.992	(0.984, 0.999)

Table 4. Spearman's rho for the correlation between community-specific relative risk estimates for accidental, respiratory-related, cardiovascular-related, and all-cause fatalities obtained using NOAA monitor-based precipitation and using NLDAS-2 Reanalysis-based data (row 1), and Spearman's rho for relative risk estimates for fatalities obtained using USGS streamflow flood data and using NOAA Storm Events data (row 2). Spearman's rho rank correlation coefficients correspond to relative risk estimates in Figure 27.

Measures	Accidental	Respiratory-related	CVD-related	All-cause
NOAA monitor-based precipitation and NLDAS-2 Reanalysis-based precipitation	0.48	0.34	0.30	0.31
USGS streamflow-based flooding and NOAA Storm Events-based flooding	-0.04	-0.01	-0.02	-0.01

FIGURES

Figure 1. 106 study locations in the contiguous United States. Honolulu, Hawaii and Anchorage, Alaska are not shown.

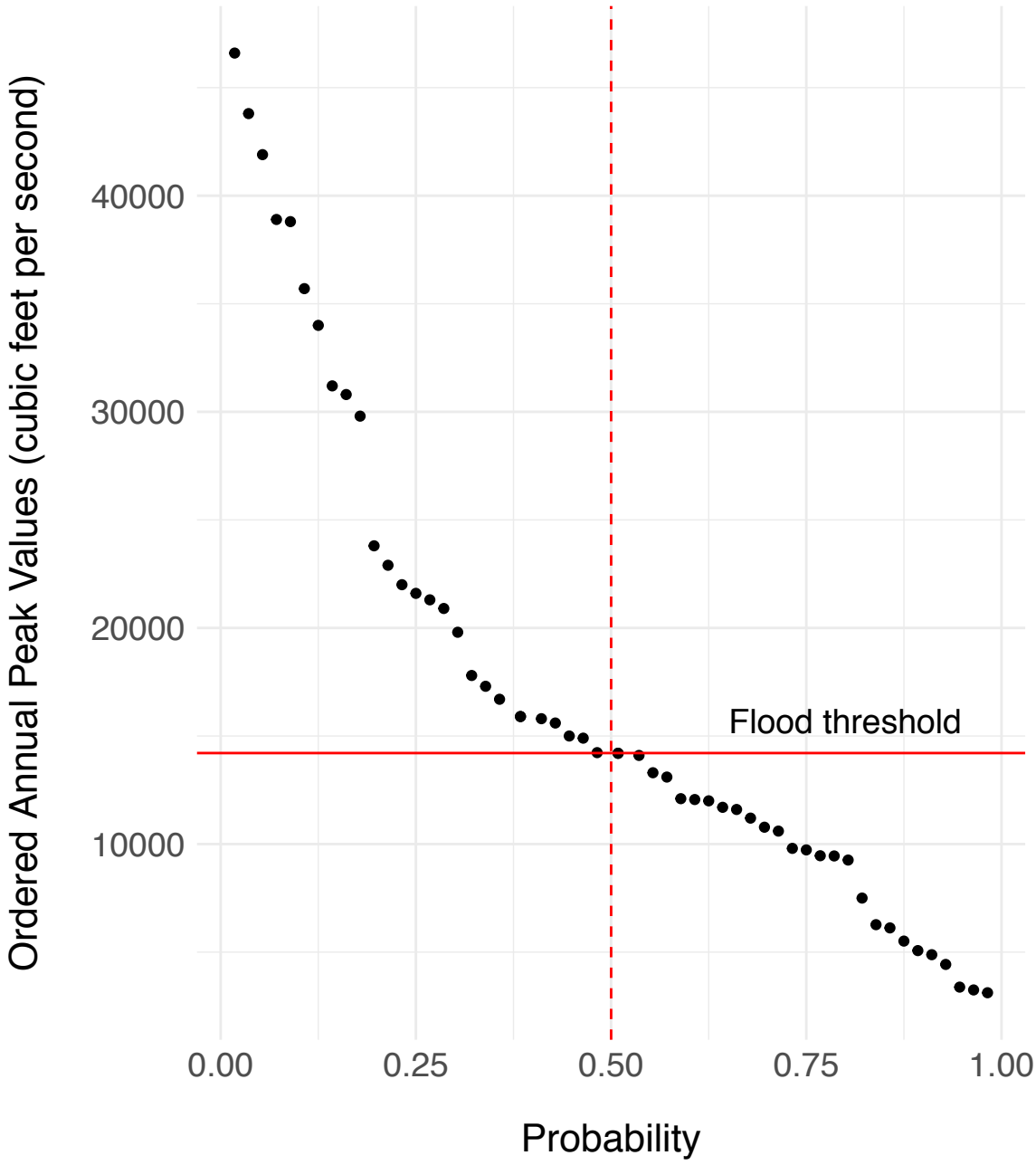


Figure 2. Example of determining a stream gage-specific median annual flood value (Q2) to use to identify flood days at the gage. The data shown is for one of 40 USGS streamflow gages available in the Los Angeles, California community, USGS streamflow gage number 11087020. Each point shows the maximum of daily streamflow values (in cubic feet per second) at the gage for one year; in total, these annual maximum streamflow values are shown for 55 years. The x-axis shows the exceedance probability for each annual maximum streamflow value. The median annual flood value (Q2), used as a flood threshold, is indicated by the horizontal red line. This value is the maximum annual streamflow that occurs at this gage with a probability of 0.5, indicated by the vertical dashed red line.

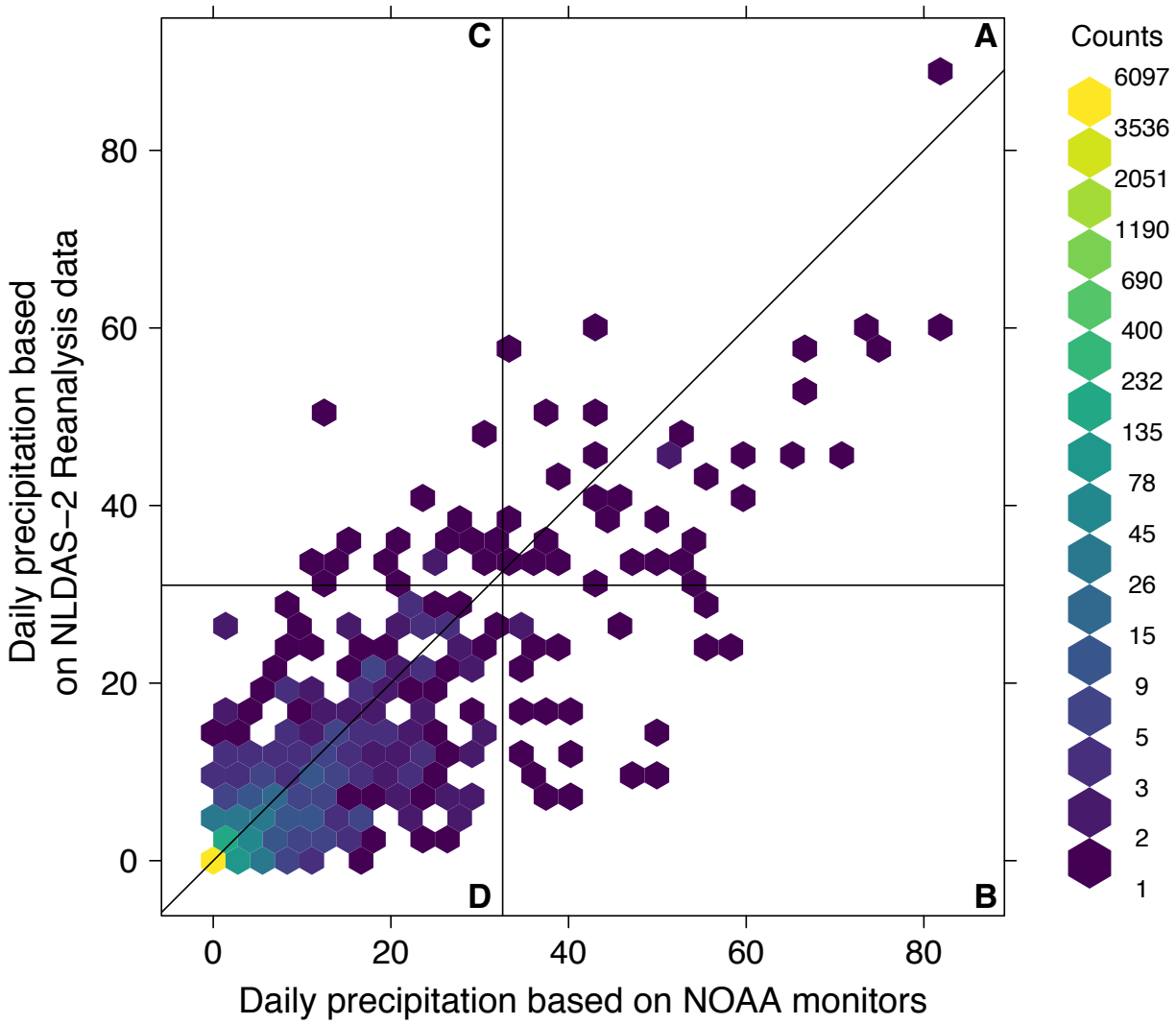


Figure 3. Two-dimensional distribution of daily precipitation values from 1987 to 2005 for two sources of precipitation data (x-axis: NOAA ground-based stations; y-axis: NLDAS-2 Reanalysis data) for Los Angeles, California. Density of study days at a location in the graph are illustrated by color, with colors closer to yellow showing a higher density of days (see legend for mapping of color to count of days over the study period). Black vertical and horizontal lines indicate the 99th percentile thresholds used to identify extreme precipitation days for the NOAA and NLDAS data (32.6 mm per day and 31.0 mm per day, respectively). A 1:1 diagonal line is included for reference and shows where values would lie in the graph if the two data sources agreed perfectly. A, B, C, and D quadrants indicate four different categories of agreement or disagreement between binary classification of extreme precipitation days based on the two precipitation data sources: days in quadrant A are classified as exposed to extreme precipitation by both data sources, days in quadrant D are classified as unexposed by both sources, and days in quadrant B and C are classified as exposed by one but not both sources.

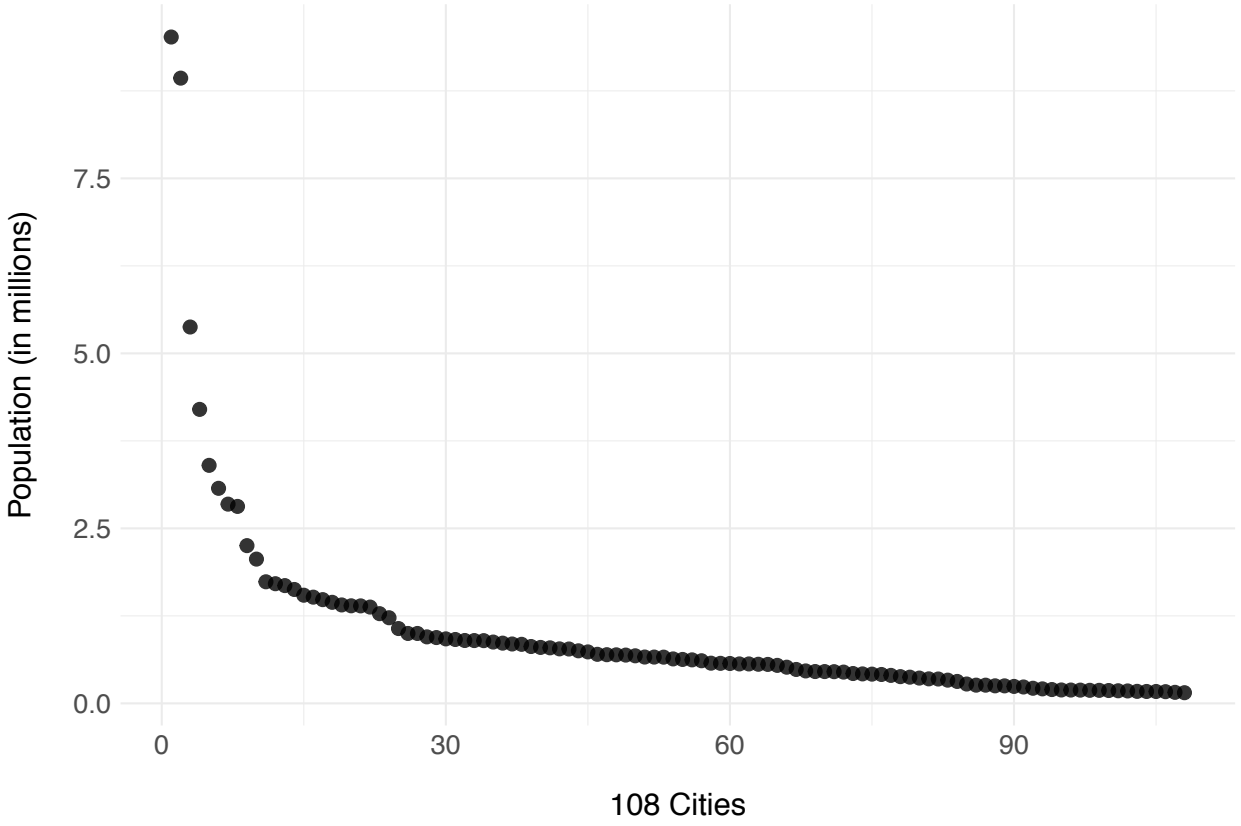


Figure 4. Populations of 108 study communities as of the 2000 U.S. Census. Each point represents one of 108 NMMAPS communities included in this study (x-axis), arranged from highest population to lowest, in millions (y-axis).

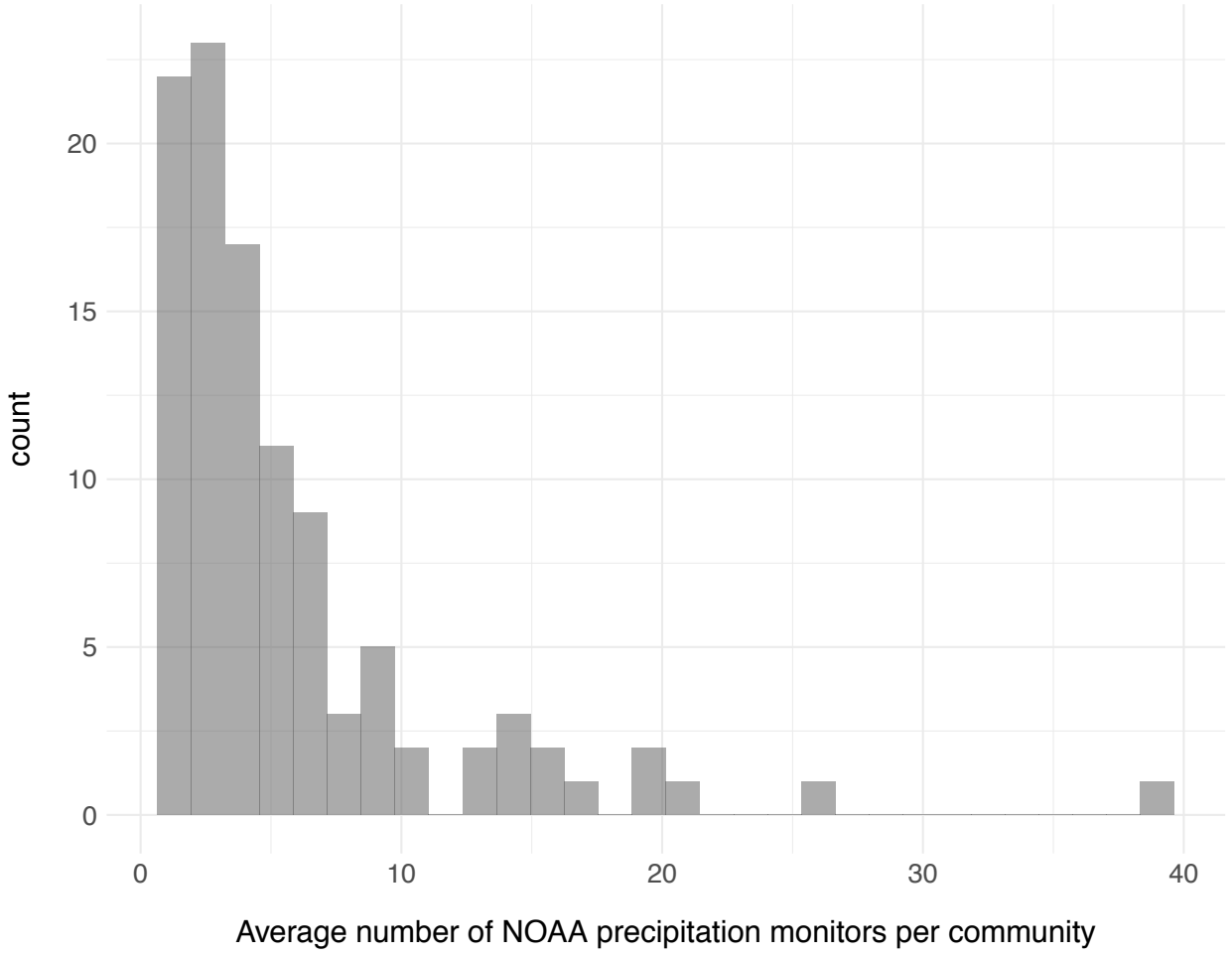


Figure 5. Histogram of the average number of stations per day available for each community's daily NOAA precipitation measurement. The x-axis gives the number of stations available in the community on average per day, while the y-axis gives the number of communities with that average.

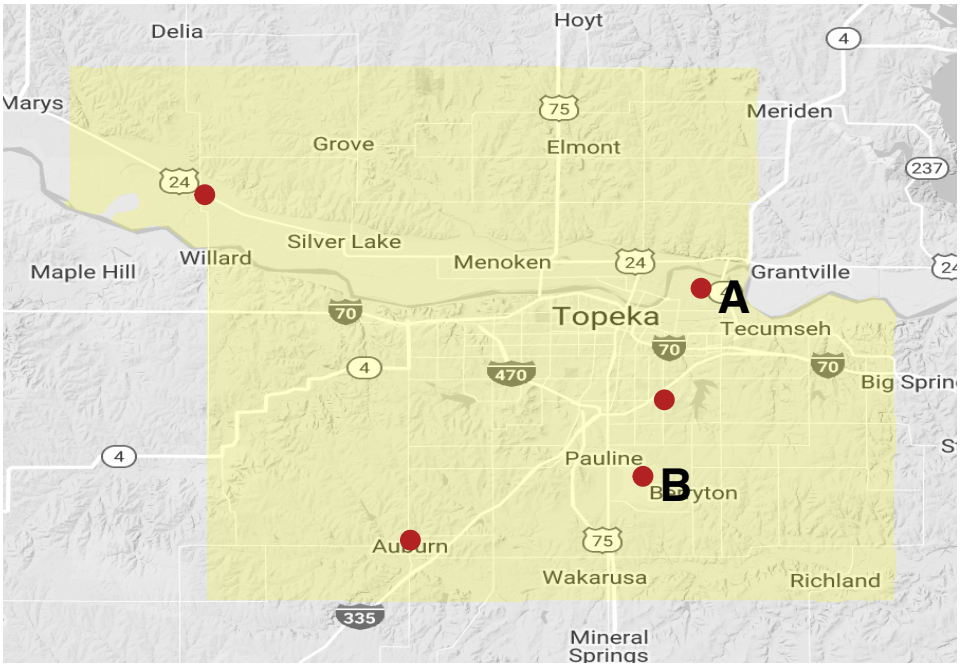


Figure 6. Locations of NOAA ground-based stations reporting precipitation values for Topeka, KS, during the study period. Yellow shading indicates the outline of the community from which data were pulled. Labels “A” and “B” indicate stations located at airports.

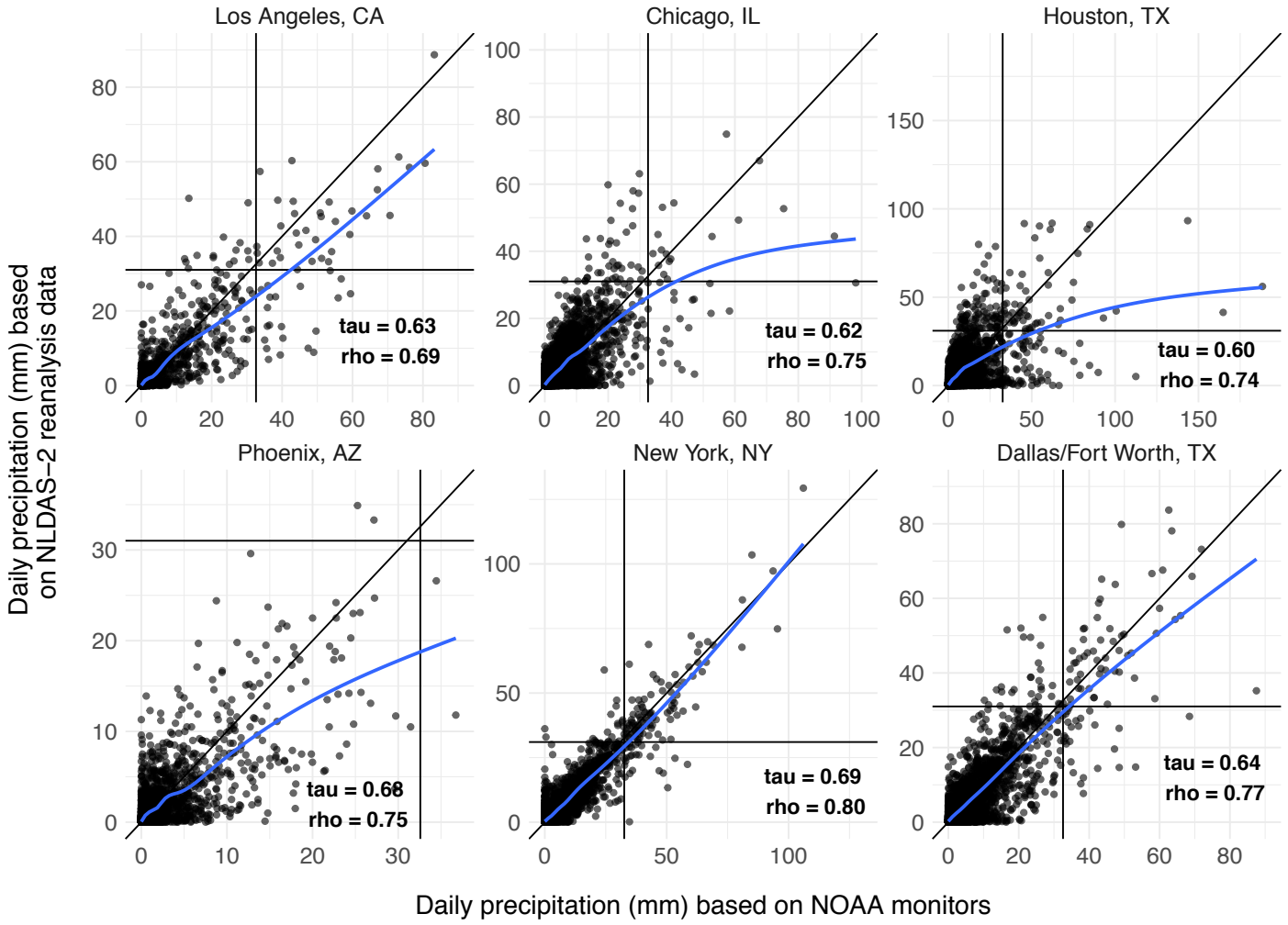


Figure 7. Daily precipitation values from 1987 to 2005 for NOAA and NLDAS data for the six largest study communities. Horizontal lines indicate 99th percentile thresholds for NOAA extreme precipitation events, and horizontal lines indicate the corresponding thresholds for NLDAS extreme precipitation events. A 1:1 line and smoothed curves based on modeling a loess smoothing function to data from NLDAS-2 regressed on data from NOAA are included for reference (blue lines).

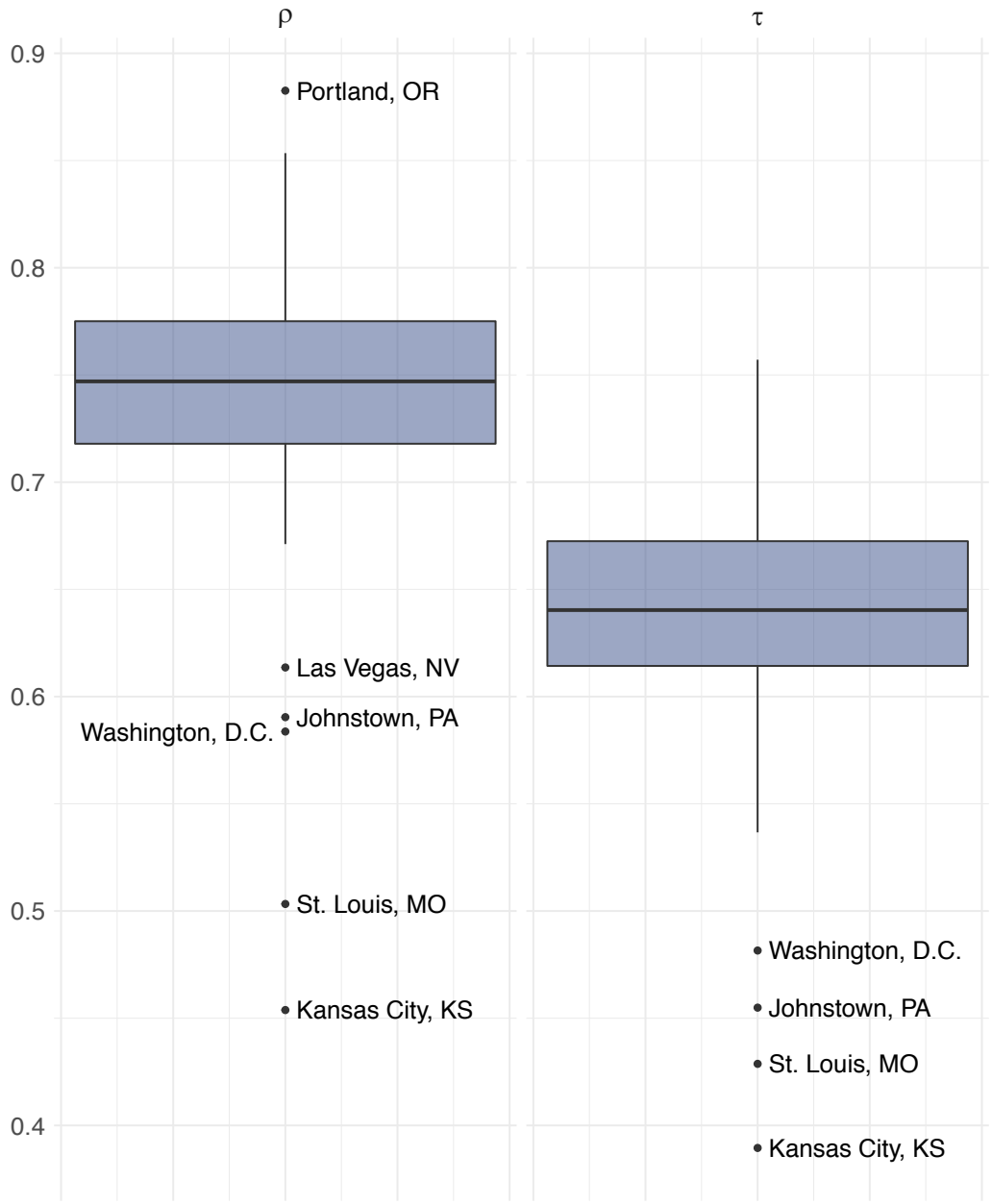


Figure 8. Boxplots of the distribution of within-community correlation between continuous daily measures of daily NOAA vs. NLDAS precipitation, as measured by Spearman's ρ (left) and Kendall's τ (right) rank correlation coefficients. The upper and lower portions of each box show the 25th and 75th percentile values of the community-level correlation coefficients, while the central line in the box shows the median value. The lines from each side of the box extend a distance of 1.5 times the inter-quartile range. Outlier communities, defined as beyond 1.5 times the interquartile range from either the 25th or 75th percentile values, are labeled.

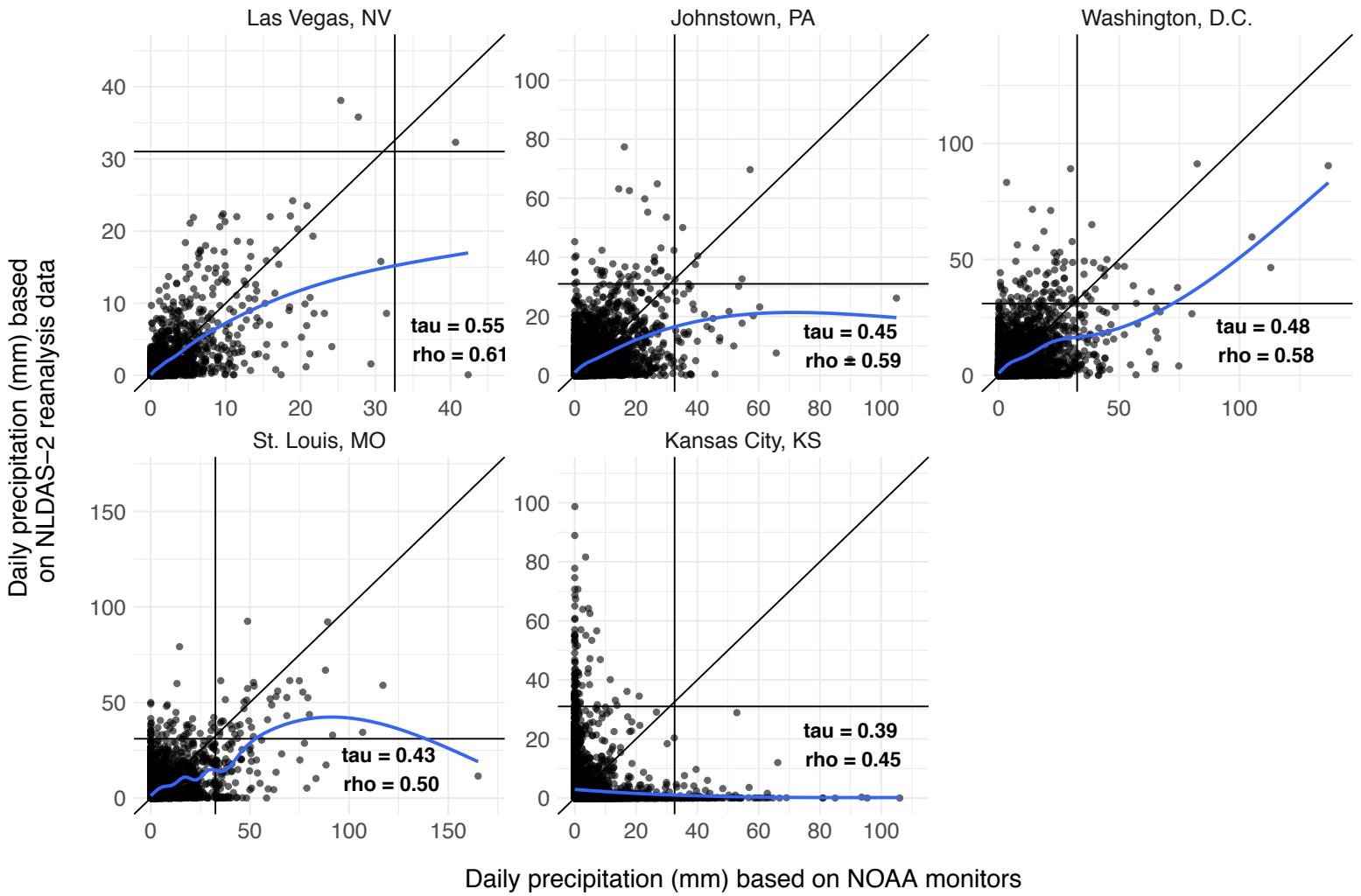


Figure 9. Daily precipitation values from 1987 to 2005 for NOAA and NLDAS data for six communities with notably low rank correlation coefficient values between continuous daily precipitation values for the two metrics. Horizontal lines indicate 99th percentile thresholds for NOAA extreme precipitation events, and horizontal lines indicate the corresponding thresholds for NLDAS extreme precipitation events. A 1:1 line and smoothed curves based on modeling a loess smoothing function to data from NLDAS-2 regressed on data from NOAA are included for reference (blue lines).

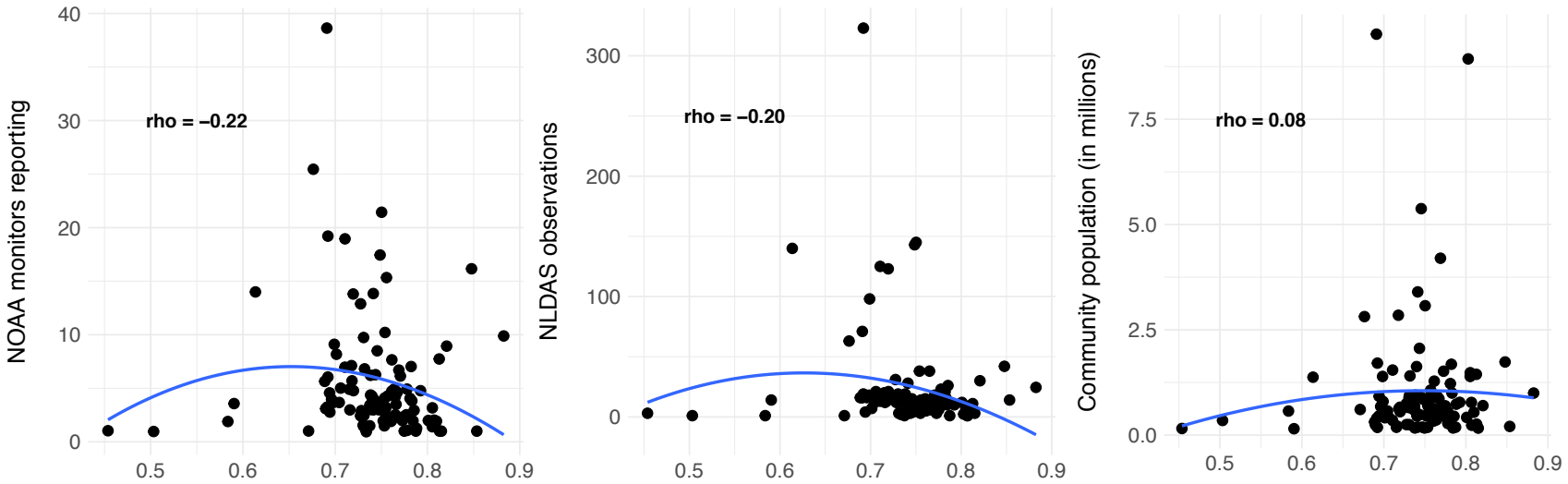


Figure 10. Scatterplots of Spearman's rho for a community (x-axis) for corrleation between NOAA and NLDAS daily precipitation values versus: average number of NOAA stations contributing to the community's precipitation value each day (left panel), average number of NLDAS observations contributing to the community's precipitation value each day (middle panel), and community population, in millions (right panel). A 1:1 line and smoothed curves based on modeling a loess smoothing function to data from NOAA stations (left panel), NLDAS observations (middle panel), and community populations (right panel) regressed on Spearman's rho are included for reference (blue lines). Spearman's rho rank correlation values are shown as text in each plot.

Figure 11. The average number of exposed days per year resulting from a nation-wide average 99th percentile extreme precipitation threshold (NOAA: 32.6 mm; NLDAS: 31.0 mm). Communities with missing data are shown with open circles, and data for Honolulu, HI, and Anchorage, AK, are not shown. There not NOAA data available for Honolulu, HI, Richmond, VA, and Newport News, VA. There was NLDAS available for all106 study communities in the contiguous United States. The distribution of the number of exposed days per year for both measures is displayed as a histogram to the right of each map. Community points are colored based on the sextile into which they fall in this exposure-specific distribution.

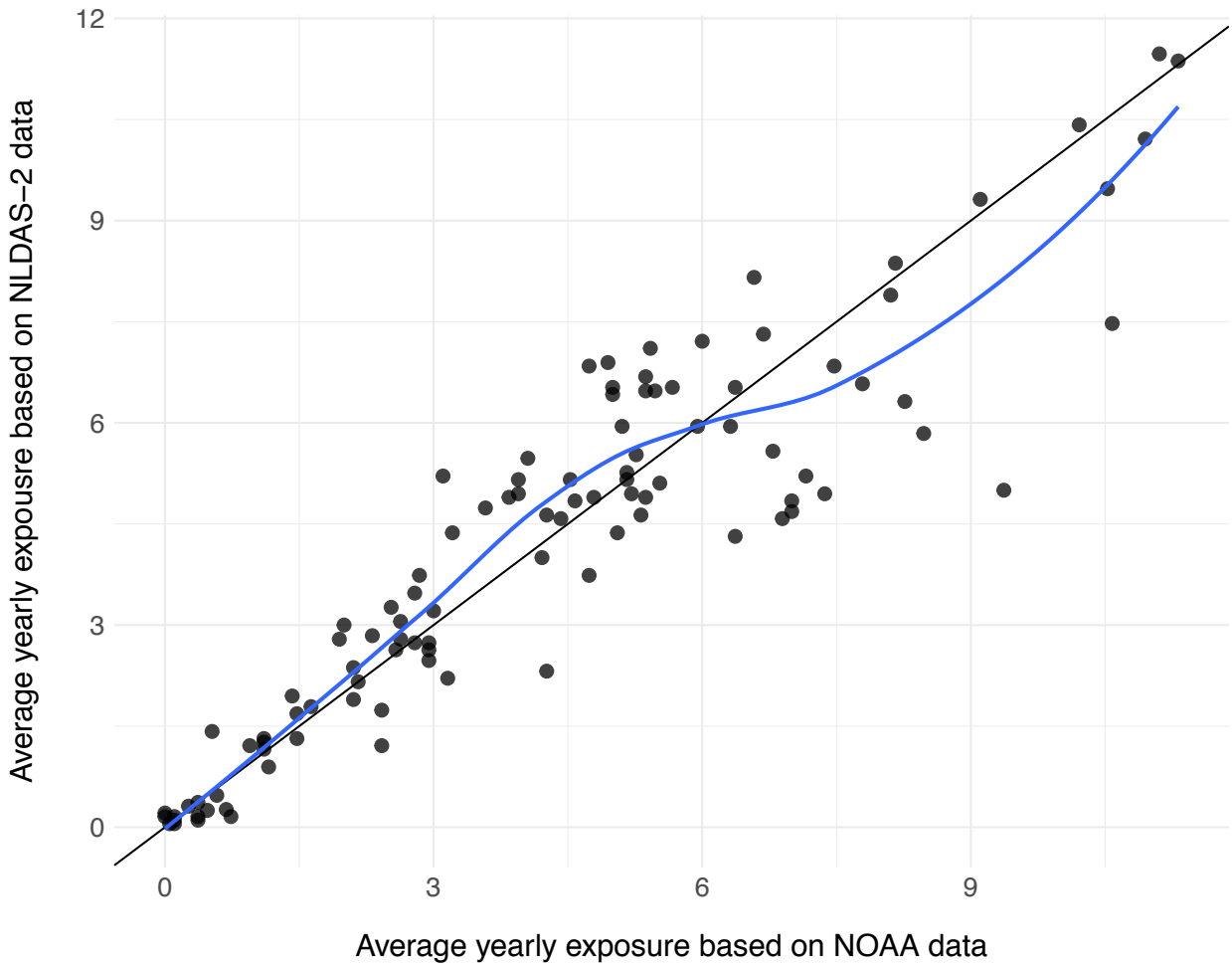


Figure 12. Association between average days of extreme precipitation exposure within a community based on the two precipitation data sources. Each point shows a study community. The x-axis shows the community's average yearly exposure days based on NOAA data and the y-axis shows the average yearly exposure based on NLDAS-2 data. A 1:1 line and smoothed curve based on modeling a loess smoothing function to data from NLDAS-2 regressed on data from NOAA are included for reference (blue line).

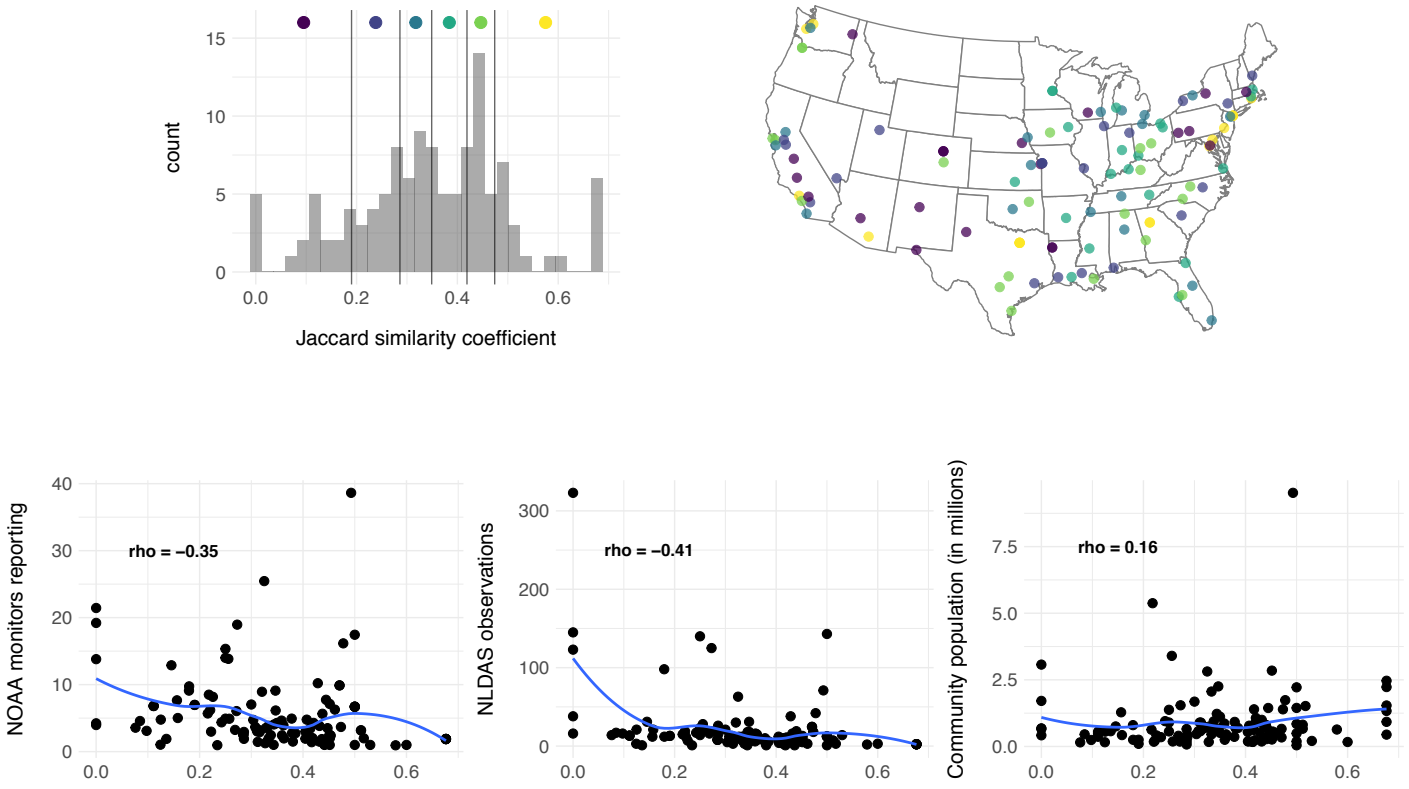


Figure 13. Histogram of Jaccard similarity coefficient values per community for NOAA monitor-based precipitation and NLDAS-2 Reanalysis precipitation data (upper left panel). Vertical lines indicate cut points for sextiles of the distribution. Colored points on the histogram correspond to colored points on the U.S. map showing geographic locations of Jaccard values (upper right panel). The lower three scatterplots show Jaccard similarity coefficient values versus: the average number of NOAA stations reporting per community, the average number of NLDAS-2 observations contributing to the average daily precipitation value per community, and each community population, in millions (lower left, lower middle, and lower right panels, respectively). Smoothed curves based on modeling a loess smoothing function to data from NOAA stations (lower left), NLDAS observations (lower middle), and community populations (lower right) regressed on Jaccard similarity coefficients are included for reference (blue lines).

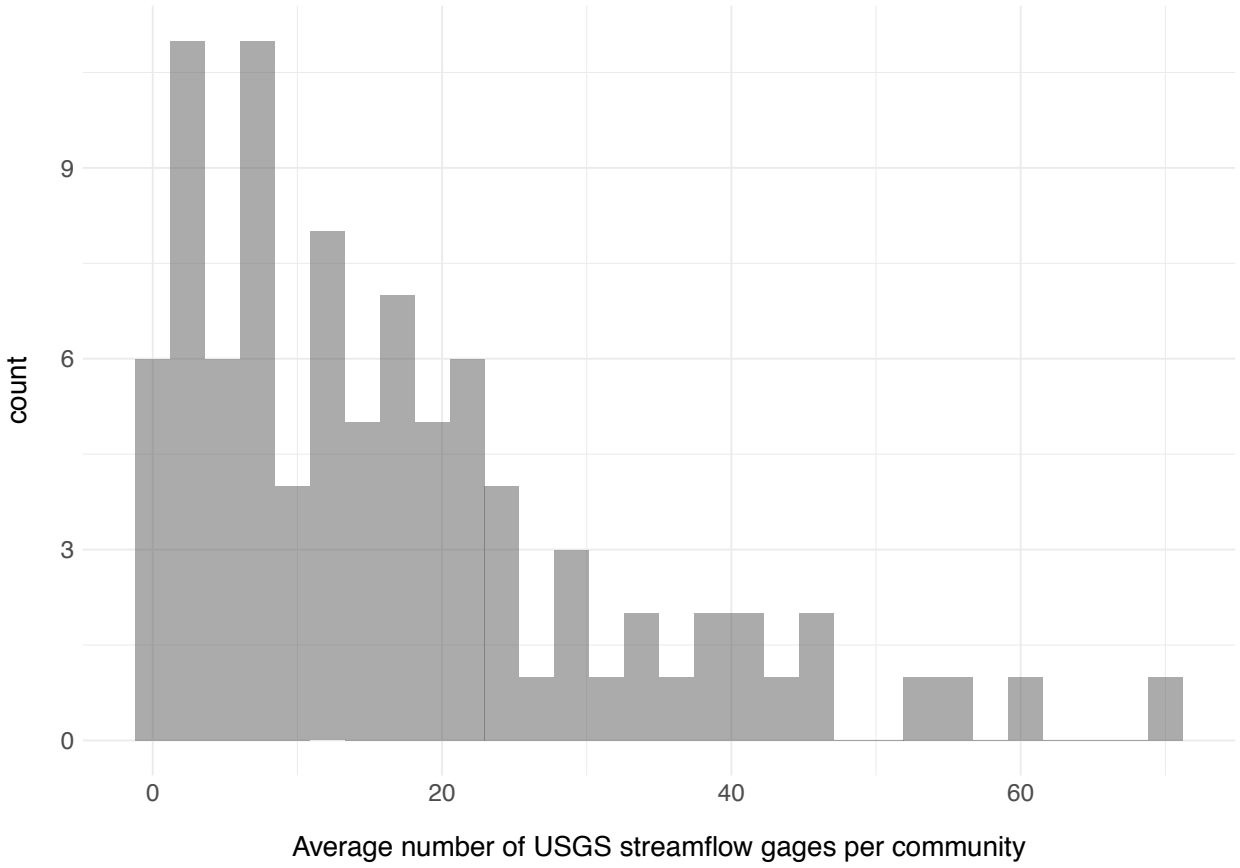


Figure 14. Histogram of the average number of streamflow gages per day (averaging only across non-missing days) in each community with streamflow data. The x-axis gives the number of USGS streamflow gages available in the community on average per day, while the y-axis gives the number of communities with that average.

Figure 15. The average number of exposed days per year for USGS and NOAA flood data. Communities with missing data are shown with open circles, and data for Honolulu, HI, and Anchorage, AK, are not shown. USGS data was unavailable for 16 communities. There was NOAA flood data available for all study communities. Due to availability of NOAA flood data, flood values from 1996 through 2005 from both sources are compared here. The distribution of exposed days per year is displayed as a histogram to the right of each map (the x-axes are displayed on a log-10 scale given the highly right-skewed distribution of these exposure estimates across communities). Community points are colored based on the sextile they fall into in this exposure-specific distribution.

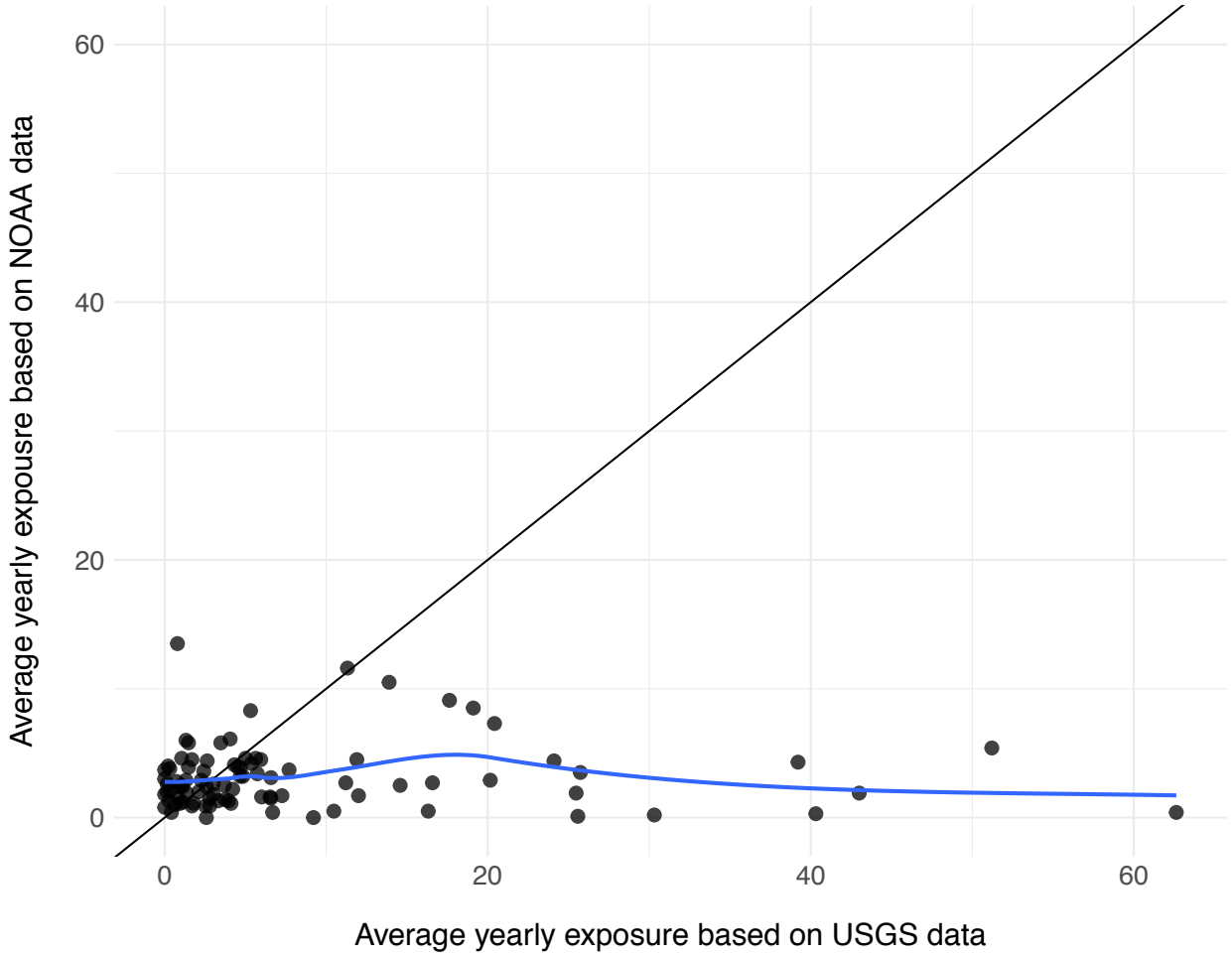


Figure 16. Association between average days of exposure to floods per year within a community based on the two flood data sources. Each point shows a study community. The x-axis shows the community's average early exposure days based on USGS data and the y-axis shows average yearly exposure based on NOAA data. A 1:1 line and smoothed curve based on modeling a loess smoothing function to data from NOAA Storm Data floods regressed on data from USGS streamflow floods are included for reference (blue line).

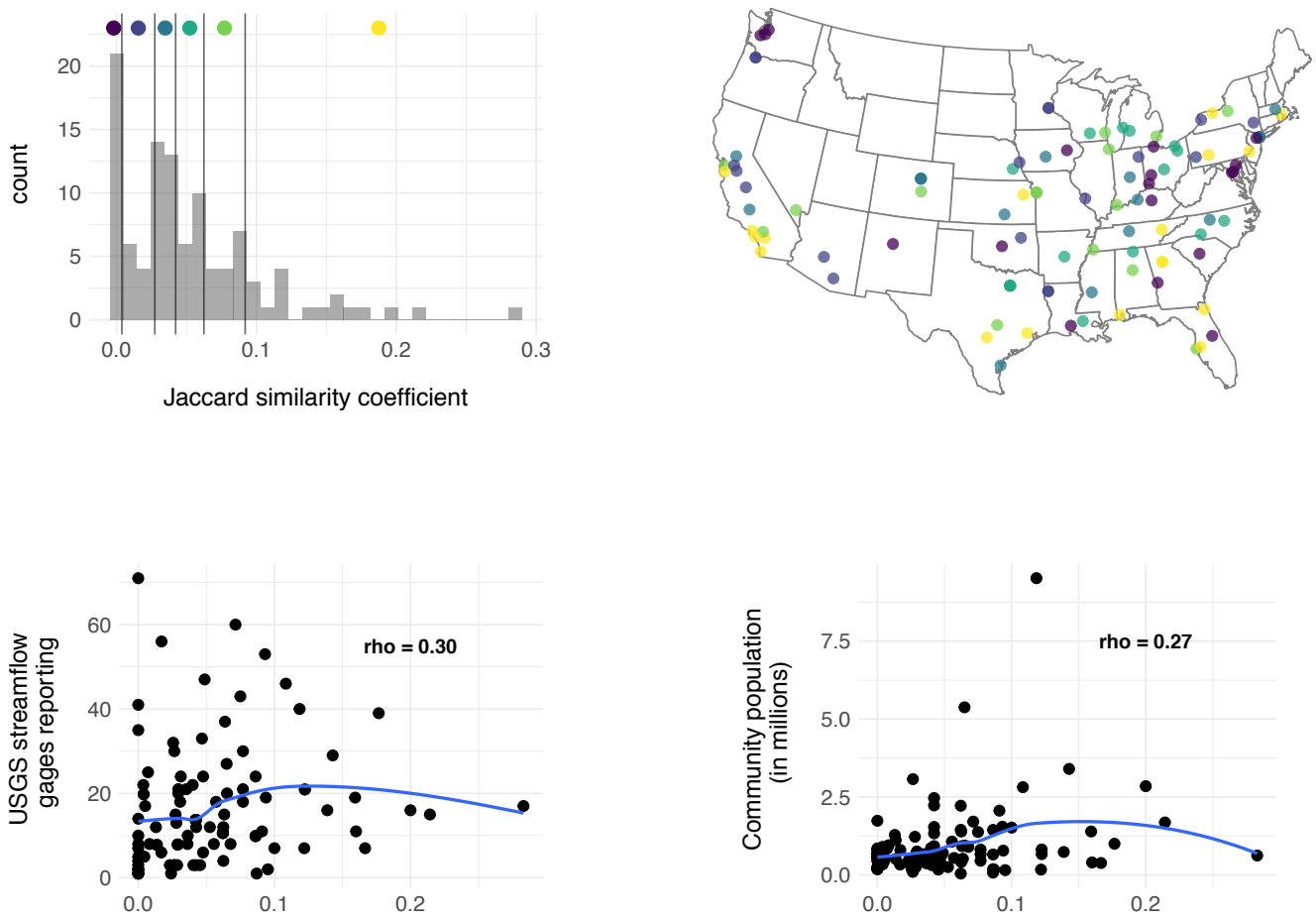


Figure 17. Histogram of Jaccard similarity coefficient values per community for USGS streamflow gage-based flooding and NOAA Storm Data-based flooding (upper left panel). Vertical lines indicate cut points for sextiles of the distribution. Colored points on the histogram correspond to colored points on the U.S. map showing geographic locations of Jaccard values (upper right panel). The lower two scatterplots show Jaccard similarity coefficient values versus: the average number of USGS streamflow gages reporting per community and each community population, in millions (lower left and lower right panels, respectively). Smoothed curves based on modeling a loess smoothing function to data from Jaccard similarity coefficients and USGS streamflow gages reporting (lower left) and community populations (lower right) are included for reference (blue lines).

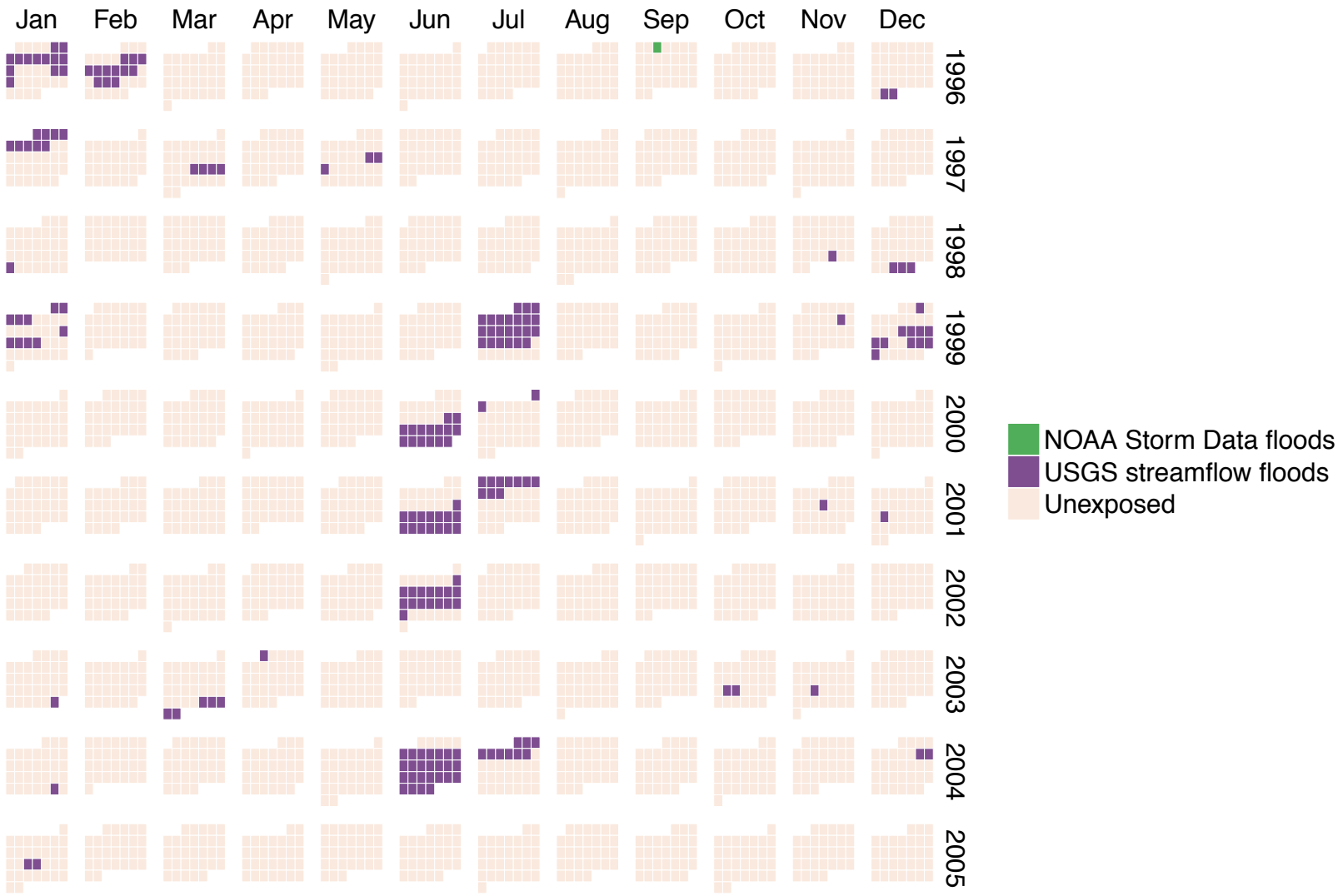


Figure 18. Calendar plot showing the number of days the Seattle, Washington community was categorized as exposed to either NOAA Storm Data floods or USGS streamflow floods from 1996 through 2005. Each colored box represents one day. Boxes are ordered by year (row), month (each block is a month within a year), and day of week (each column within a block).

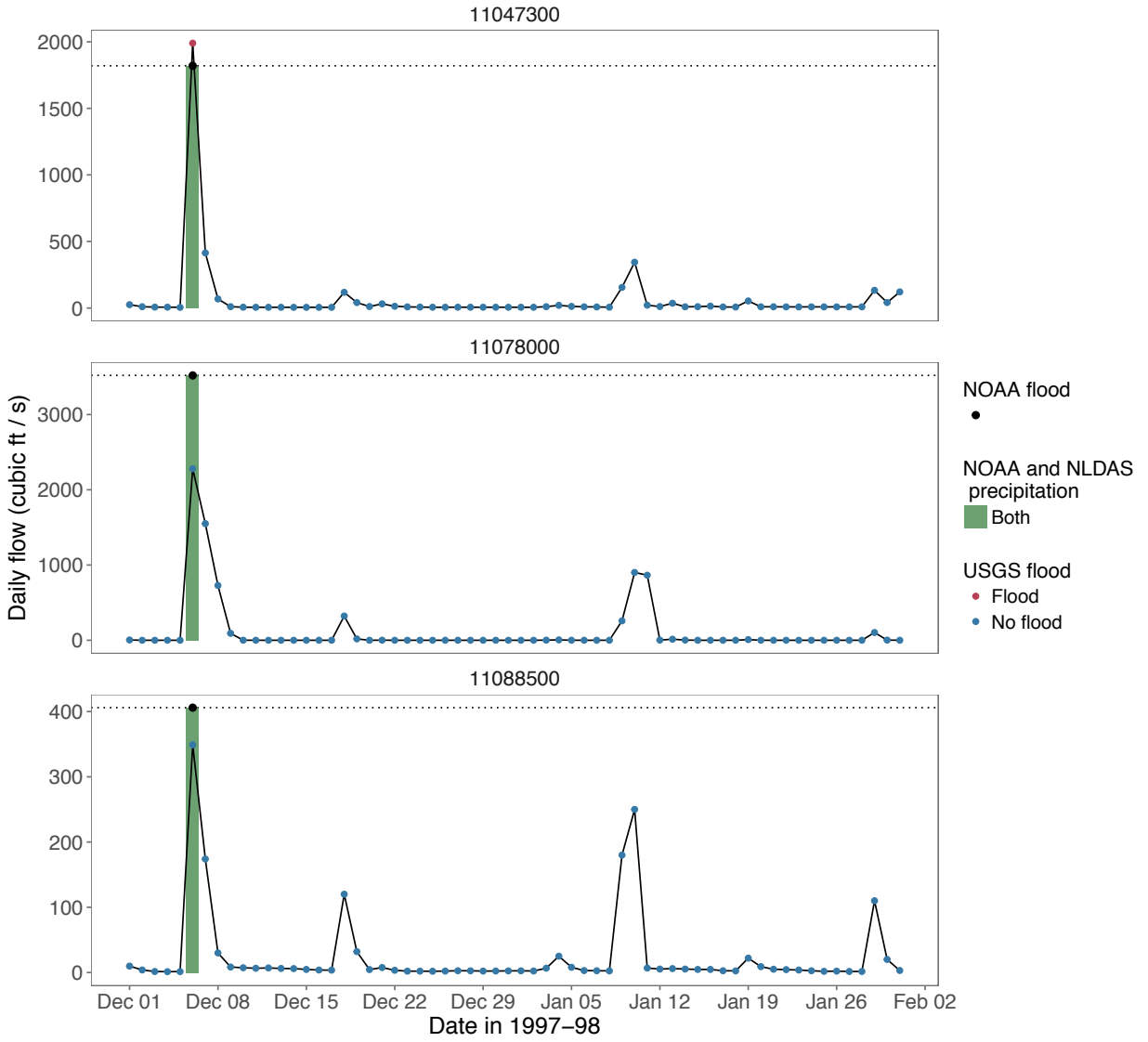


Figure 19. USGS streamflow and flood events, NOAA flood events, and NOAA and NLDAS extreme precipitation events in the Santa Ana/Anaheim community in California from December 1, 1997 through February 1, 1998. Each facet of the figure shows the streamflow from one of the USGS streamflow gages contributing to the average flood measure for the community. Streamflow is measured in cubic feet per second. Horizontal dotted lines represent gage-specific flood thresholds. Blue points indicate days on which there was not a USGS flood, and red points indicate days on which there was a USGS flood recorded. Black points falling on the USGS flood threshold line represent days for which a flood was recorded in the NOAA Storm Event database. Green columns represent days for which both NOAA and NLDAS data recorded an extreme precipitation day. There were no days in this period for which only NOAA or only NLDAS data indicated an extreme precipitation day.

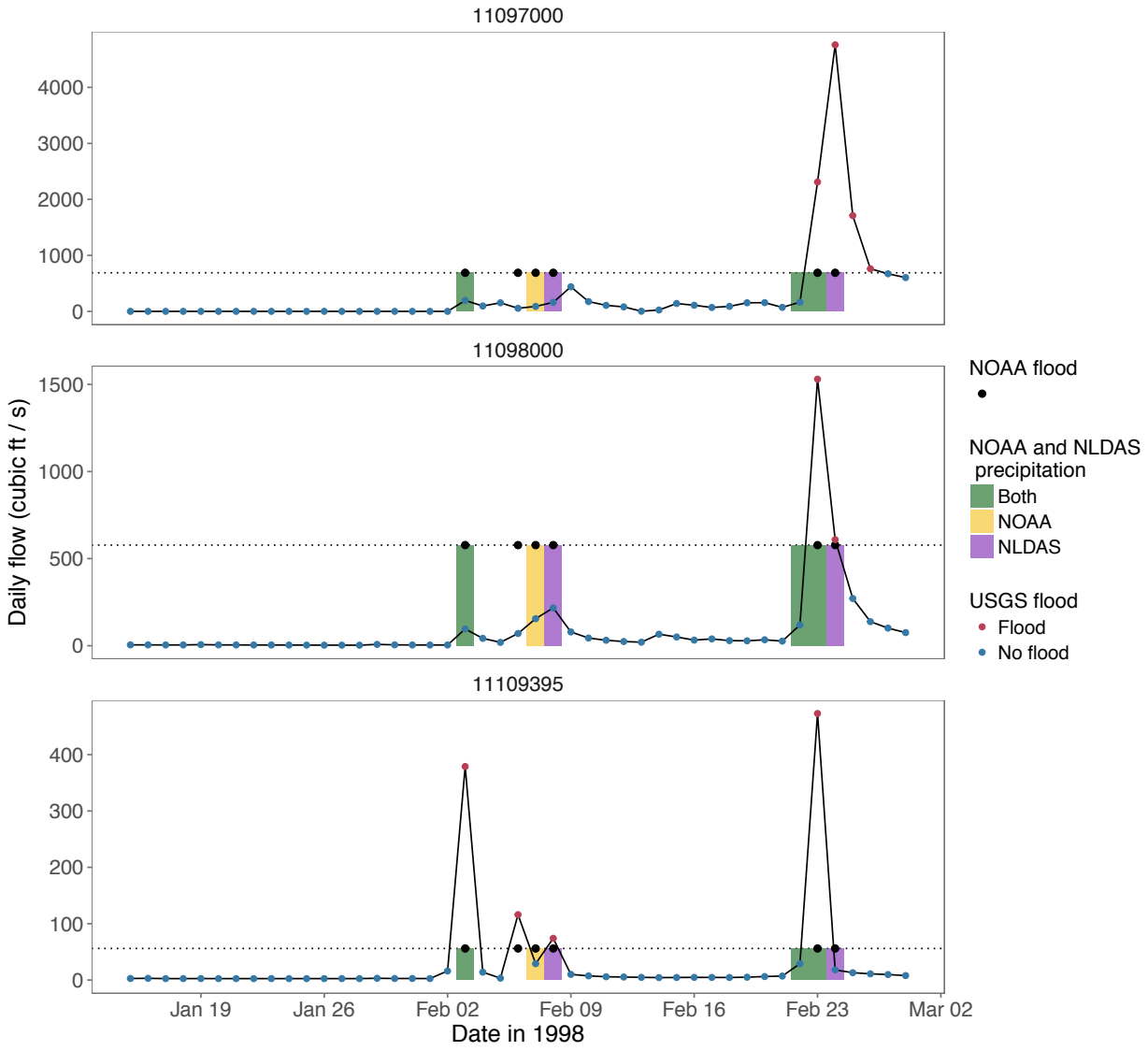


Figure 20. USGS streamflow and flood events, NOAA flood events, and NOAA and NLDAS extreme precipitation events in the Santa Ana/Anaheim community in California from December 1, 1997 through February 1, 1998. Each facet of the figure shows the streamflow from one of the USGS streamflow gages contributing to the average flood measure for the community. Streamflow is measured in cubic feet per second. Horizontal dotted lines represent gage-specific flood thresholds. Black points falling on the USGS flood threshold line represent days for which a flood was recorded in the NOAA Storm Event database. Green columns represent days for which both NOAA and NLDAS data recorded an extreme precipitation day, yellow columns represent days on which only NOAA recorded extreme precipitation, and purple columns represent days on which only NLDAS recorded extreme precipitation.

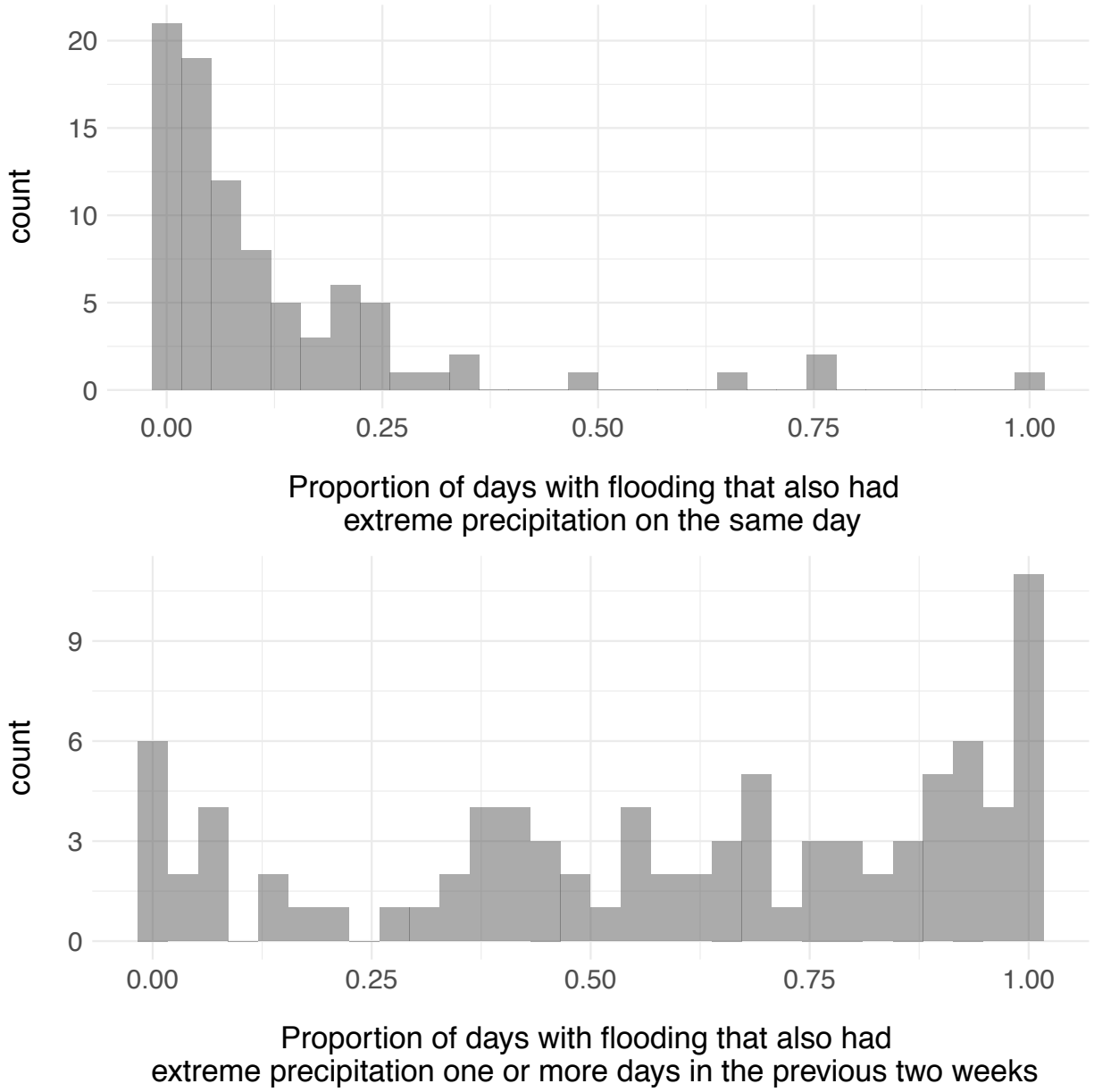


Figure 21. Top panel: histogram of the proportion of days with extreme flooding using USGS data and extreme rain using NLDAS-2 data on the same day in each community with NLDAS-2 and USGS data. Bottom panel: histogram of the proportion of days with flooding and extreme precipitation on the same day or at least one day in the previous 14 days. The x-axis gives the proportion *P* for each community, and the y-axis gives the number of communities with that proportion.

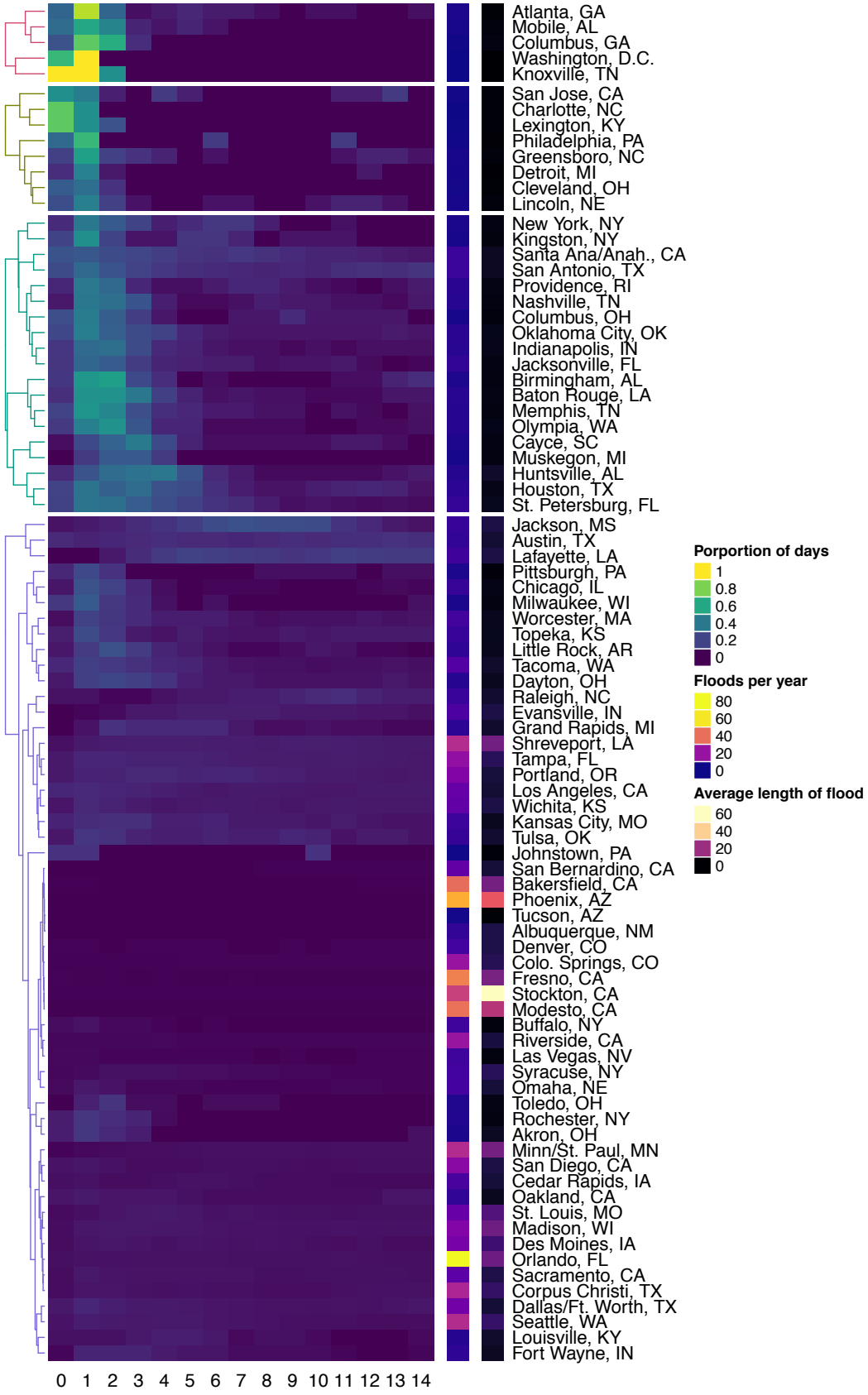

Atlanta, GA
Mobile, AL
Columbus, GA
Washington, D.C.
Knoxville, TN
San Jose, CA
Charlotte, NC
Lexington, KY
Philadelphia, PA
Greensboro, NC
Detroit, MI
Cleveland, OH
Lincoln, NE
New York, NY
Kingston, NY
Santa Ana/Anah., CA
San Antonio, TX
Providence, RI
Nashville, TN
Columbus, OH
Oklahoma City, OK
Indianapolis, IN
Jacksonville, FL
Birmingham, AL
Baton Rouge, LA
Memphis, TN
Olympia, WA
Cayce, SC
Muskegon, MI
Huntsville, AL
Houston, TX
St. Petersburg, FL
Jackson, MS
Austin, TX
Lafayette, LA
Pittsburgh, PA
Chicago, IL
Milwaukee, WI
Worcester, MA
Topeka, KS
Little Rock, AR
Tacoma, WA
Dayton, OH
Raleigh, NC
Evansville, IN
Grand Rapids, MI
Shreveport, LA
Tampa, FL
Portland, OR
Los Angeles, CA
Wichita, KS
Kansas City, MO
Tulsa, OK
Johnstown, PA
San Bernardino, CA
Bakersfield, CA
Phoenix, AZ
Tucson, AZ
Albuquerque, NM
Denver, CO
Colo. Springs, CO
Fresno, CA
Stockton, CA
Modesto, CA
Buffalo, NY
Riverside, CA
Las Vegas, NV
Syracuse, NY
Omaha, NE
Toledo, OH
Rochester, NY
Akron, OH
Minn/St. Paul, MN
San Diego, CA
Cedar Rapids, IA
Oakland, CA
St. Louis, MO
Madison, WI
Des Moines, IA
Orlando, FL
Sacramento, CA
Corpus Christi, TX
Dallas/Ft. Worth, TX
Seattle, WA
Louisville, KY
Fort Wayne, IN
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Porportion of days
1
0.8
0.6
0.4
0.2
0
Floods per year
80
60
40
20
0
Average length of flood
60
40
20
0

Figure 22. Heat map showing the distribution of proportions *P* of days with flooding that also had extreme precipitation on a day a certain lag before the flood day, as measured with NLDAS precipitation data, from the day of the flood up to two weeks prior to the flood day. Columns in the main heat map indicate the lag day. The "0" column indicates values of *P* for same-day comparisons, while the "1" column, for example, indicates values for *P* comparing days with flooding with extreme precipitation events in the previous day. Values for the average number of USGS floods per year and the average length of flood (measured in days) are included in columns to the right of the main *P* heat map. Communities are hierarchically clustered into four groups, indicated by row dendrograms to the left of the heat map.

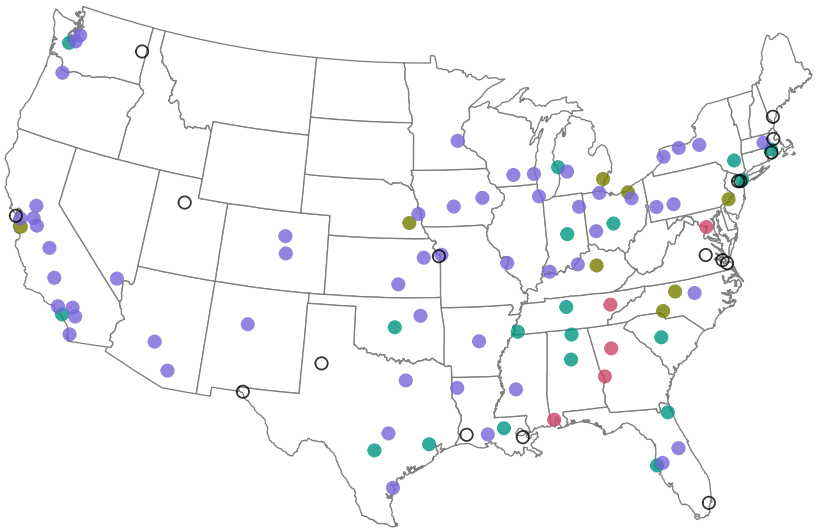


Figure 23. Geographic locations of communities included in the NLDAS precipitation vs. USGS flooding proportion *P* calculation. Open circles indicate communities without available USGS streamflow data. Point colors correspond to the colors of the row dendrograms in Figure 16.

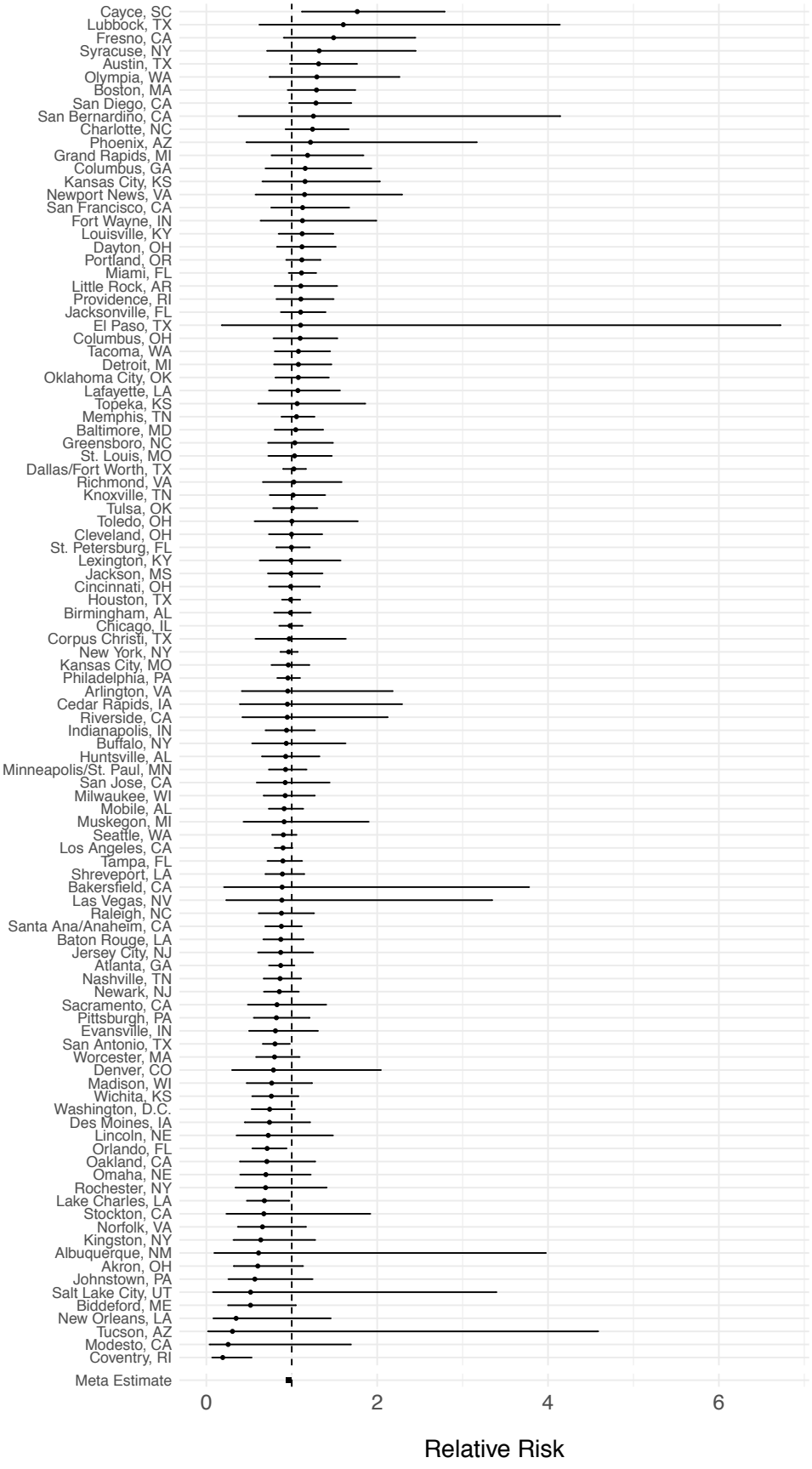

Cayce, SC
Lubbock, TX
Fresno, CA
Syracuse, NY
Austin, TX
Olympia, WA
Boston, MA
San Diego, CA
San Bernardino, CA
Charlotte, NC
Phoenix, AZ
Grand Rapids, MI
Columbus, GA
Kansas City, KS
Newport News, VA
San Francisco, CA
Fort Wayne, IN
Louisville, KY
Dayton, OH
Portland, OR
Miami, FL
Little Rock, AR
Providence, RI
Jacksonville, FL
El Paso, TX
Columbus, OH
Tacoma, WA
Detroit, MI
Oklahoma City, OK
Lafayette, LA
Topeka, KS
Memphis, TN
Baltimore, MD
Greensboro, NC
St. Louis, MO
Dallas/Fort Worth, TX
Richmond, VA
Knoxville, TN
Tulsa, OK
Toledo, OH
Cleveland, OH
St. Petersburg, FL
Lexington, KY
Jackson, MS
Cincinnati, OH
Houston, TX
Birmingham, AL
Chicago, IL
Corpus Christi, TX
New York, NY
Kansas City, MO
Philadelphia, PA
Arlington, VA
Cedar Rapids, IA
Riverside, CA
Indianapolis, IN
Buffalo, NY
Huntsville, AL
Minneapolis/St. Paul, MN
San Jose, CA
Milwaukee, WI
Mobile, AL
Muskegon, MI
Seattle, WA
Los Angeles, CA
Tampa, FL
Shreveport, LA
Bakersfield, CA
Las Vegas, NV
Raleigh, NC
Santa Ana/Anaheim, CA
Baton Rouge, LA
Jersey City, NJ
Atlanta, GA
Nashville, TN
Newark, NJ
Sacramento, CA
Pittsburgh, PA
Evansville, IN
San Antonio, TX
Worcester, MA
Denver, CO
Madison, WI
Wichita, KS
Washington, D.C.
Des Moines, IA
Lincoln, NE
Orlando, FL
Oakland, CA
Omaha, NE
Rochester, NY
Lake Charles, LA
Stockton, CA
Norfolk, VA
Kingston, NY
Albuquerque, NM
Akron, OH
Johnstown, PA
Salt Lake City, UT
Biddeford, ME
New Orleans, LA
Tucson, AZ
Modesto, CA
Coventry, RI
Meta Estimate
0
2
4
6
Relative Risk

Figure 24. Community-specific effect estimates of the association between extreme precipitation (determined using NLDAS precipitation data) and risk of accidental mortality. Models were adjusted for temperature, the day of the week, and long-term seasonal trends. Each point shows the effect estimate for a specific community and horizontal line shows the 95% confidence interval estimated for the community. Communities are ordered by their central point estimates of association between extreme precipitation exposure and accidental mortality risk. The pooled association estimate across all communities is shown at the bottom of plot ("Meta Estimate").

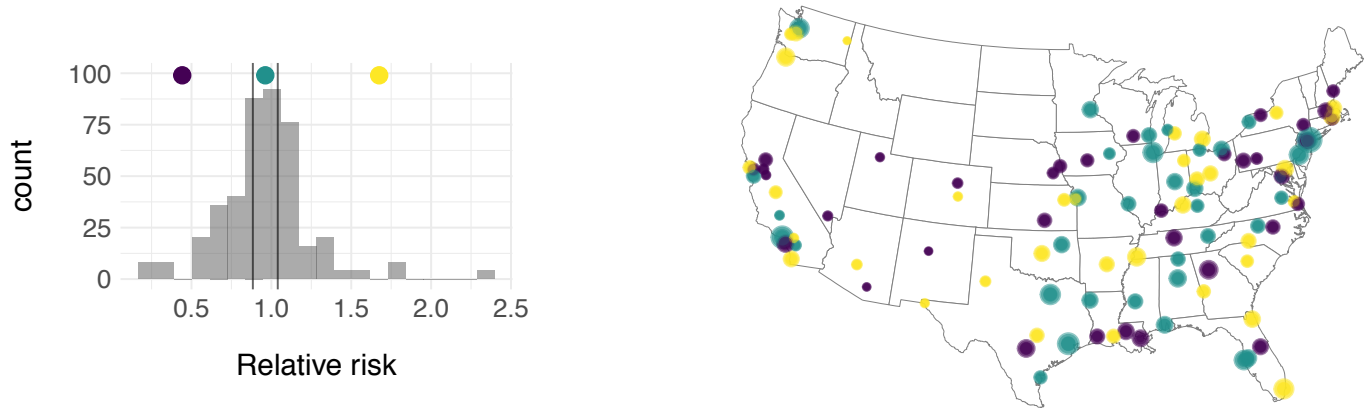


Figure 25. Histogram of relative risk estimates for the association between extreme rainfall (determined using NLDAS precipitation data) and risk of accidental mortality (left). Vertical lines indicate cut points for quartiles of the distribution. Colored points on the histogram correspond to colored points on the U.S. map showing geographic locations of relative risk estimates for each community (right). The size of points on the map corresponds to the inverse of the standard error of relative risk estimates.

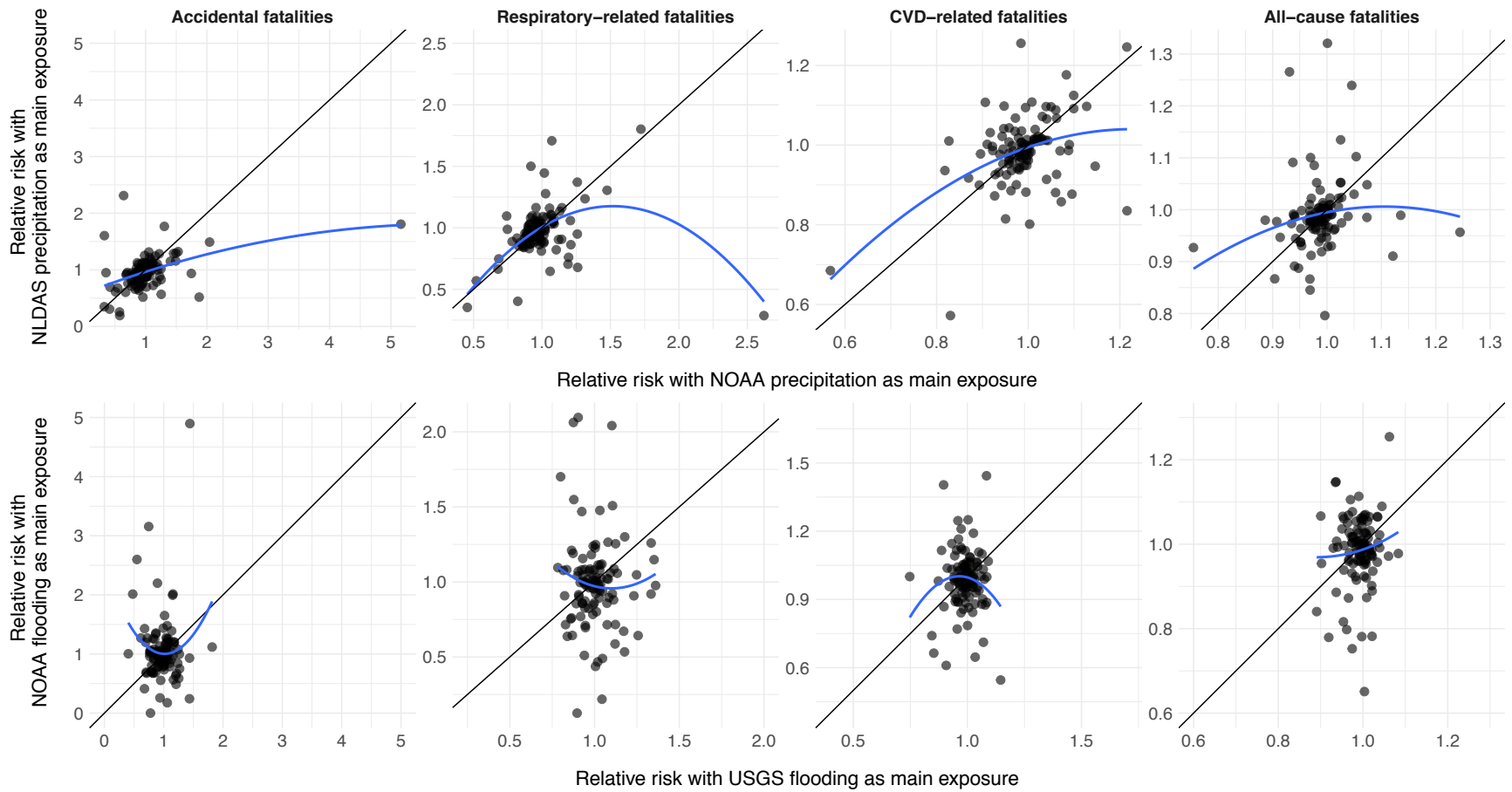


Figure 26. Scatterplots showing the relative risk estimates for accidental, respiratory-related, cardiovascular-related, and all-cause fatalities (left, middle-left, middle-right, and right panels, respectively) for NOAA monitor-based precipitation versus NLDAS-2 Reanalysis-based precipitation data (upper panel) and for USGS streamflow gage-based flooding versus NOAA Storm Events-based flooding (lower panel). A 1:1 line is included for reference in each panel, as well as smooth curves based on modeling a loess smoothing function to data from relative risks with NLDAS precipitation (upper row) or NOAA flooding (lower row) regressed on relative risks with NOAA precipitation (upper row) or USGS flooding (lower row) (blue lines).

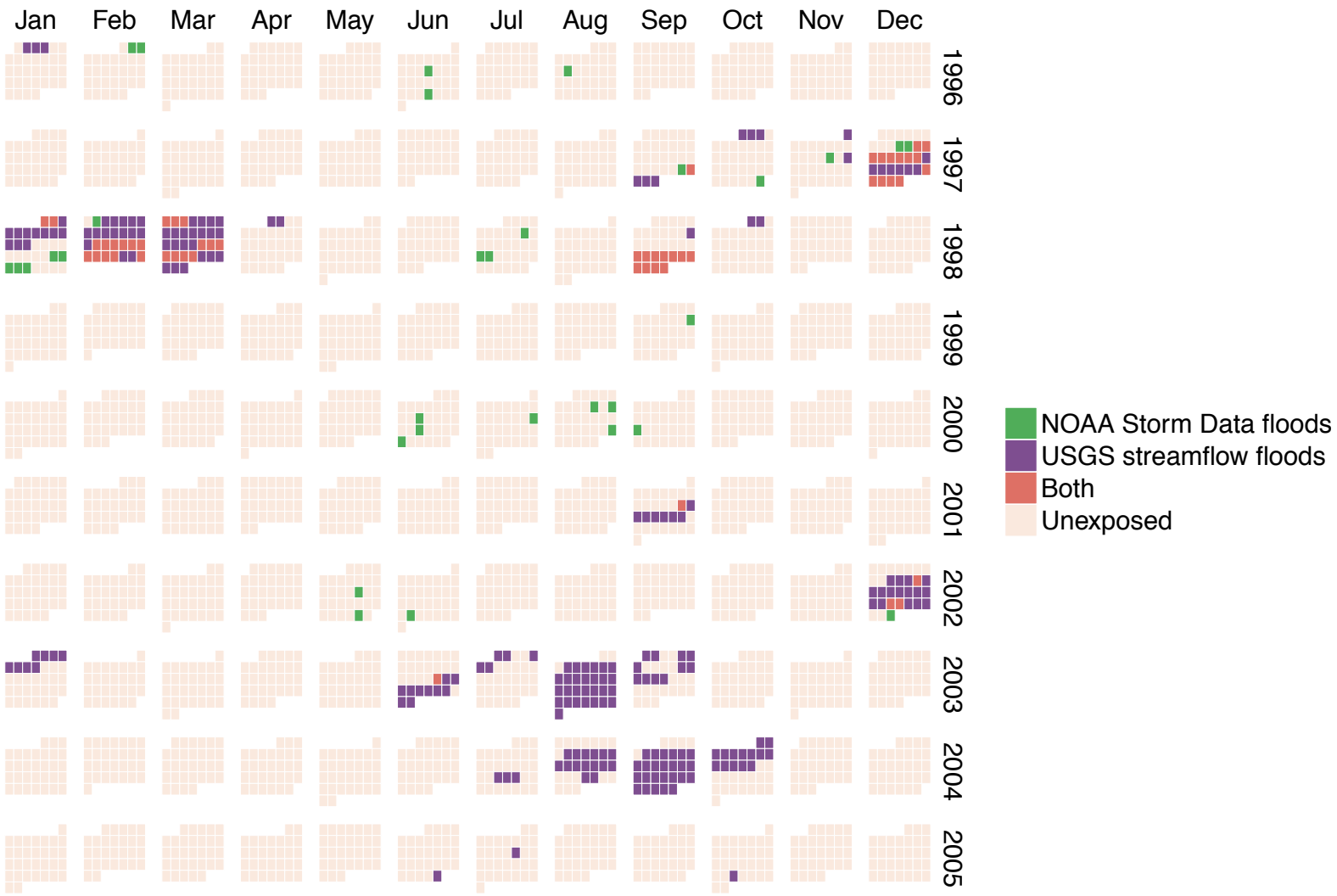


Figure 27. Calendar plot showing the number of days the Tampa, FL, community was categorized as exposed to either NOAA Storm Data floods, USGS streamflow floods, or both from 1996 through 2005. Each colored box represents one day. Boxes are ordered by year (row), month (each block is a month within a year), and day of week (each column within a block).

REFERENCES

Ahern, Mike, R. Sari Kovats, Paul Wilkinson, Roger Few, and Franziska Matthies. 2005. "Global Health Impacts of Floods: Epidemiologic Evidence." *Epidemiologic Reviews* 27: 36–46.

Alderman, Katarzyna, Lyle R. Turner, and Shilu Tong. 2012. "Floods and Human Health: A Systematic Review." *Environment International* 47. Elsevier B.V.: 37–47.

Anderson, Brooke G, and Michelle L Bell. 2009. "Weather-Related Mortality: How Heat, Cold, and Heat Waves Affect Mortality in the United States." *Epidemiology (Cambridge, Mass.)* 20 (2): 205–13.

Anderson, G. Brooke, Francesca Dominici, Yun Wang, Meredith C. McCormack, Michelle L. Bell, and Roger D. Peng. 2013. "Heat-Related Emergency Hospitalizations for Respiratory Diseases in the Medicare Population." *American Journal of Respiratory and Critical Care Medicine* 187 (10): 1098–1103.

Anderson, G Brooke, and Michelle L Bell. 2011. "Heat Waves in the United States: Mortality Risk during Heat Waves and Effect Modification." *Environmental Health Perspectives* 119 (2): 210–18.

Anderson, W, G J Prescott, S Packham, J Mullins, M Brookes, and a Seaton. 2001. "Asthma Admissions and Thunderstorms: A Study of Pollen, Fungal Spores, Rainfall, and Ozone." *QJM : Monthly Journal of the Association of Physicians* 94: 429–33.

Ashley, W S, S Strader, D C Dziubla, and A Haberlie. 2015. "DRIVING BLIND Weather-Related Vision Hazards and Fatal Motor Vehicle Crashes." *Bulletin of the American Meteorological Society* 96 (May): 755–78.

Barnett, Adrian G., Cunrui Huang, and Lyle Turner. 2012. "Benefits of Publicly Available

Data." *Epidemiology* 23 (3): 500–501.

Bell, J.E., S.C. Herring, L. Jantarasami, C. Adrianopoli, K. Benedict, K. Conlon, V. Escobar, J. Hess, J. Luvall, C.P. Garcia-Pando, D. Quattrochi, J. Runkle, and C.J. Schreck, III, 2016: Ch. 4: Impacts of Extreme Events on Human Health. The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment. U.S. Global Change Research Program, Washington, DC, 99–128.

Bell, Michelle L, Aidan Mcdermott, Scott L Zeger, and Jonathan M Samet. 2004. "In 95 US Urban Communities , 1987-2000." *Forestry* 292 (19): 2372–78.

Bergel-Hayat, Ruth, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. 2013. "Explaining the Road Accident Risk: Weather Effects." *Accident Analysis and Prevention* 60. Elsevier Ltd: 456–65.

Black, Alan W., Gabriele Villarini, and Thomas L. Mote. 2017. "Effects of Rainfall on Vehicle Crashes in Six U.S. States." *Weather, Climate, and Society* 9 (1): 53–70.

Borga, Marco, Markus Stoffel, Lorenzo Marchi, Francesco Marra, and Matthias Jakob. 2014. "Hydrogeomorphic Response to Extreme Rainfall in Headwater Systems: Flash Floods and Debris Flows." *Journal of Hydrology* 518 (PB). Elsevier B.V.: 194–205.

Brandsema, P S, S M Euser, I Karagiannis, J W DEN Boer, and W VAN DER Hoek. 2014. "Summer Increase of Legionnaires' Disease 2010 in The Netherlands Associated with Weather Conditions and Implications for Source Finding." *Epidemiology and Infection*, 1–12.

Brook, Robert D., Sanjay Rajagopalan, C. Arden Pope, Jeffrey R. Brook, Aruni Bhatnagar, Ana V. Diez-Roux, Fernando Holguin, et al. 2010. "Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart

Association." *Circulation* 121 (21): 2331–78.

Browne, Mark J., and Robert E. Hoyt. 1999. "The Demand for Flood Insurance: Empirical Evidence." *SSRN Electronic Journal*, 291–306.

Brun, S. E., and L. E. Band. 2000. "Simulating Runoff Behavior in an Urbanizing Watershed." *Computers, Environment and Urban Systems* 24 (1): 5–22.

Brunekreef, Bert, and Stephen T. Holgate. 2002. "Air Pollution and Health." *Lancet* 360 (9341): 1233–42.

Buguet, Alain. 2007. "Sleep under Extreme Environments: Effects of Heat and Cold Exposure, Altitude, Hyperbaric Pressure and Microgravity in Space." *Journal of the Neurological Sciences* 262 (1–2): 145–52.

Campese, Christine, Dounia Bitar, Sophie Jarraud, Catherine Maine, Françoise Forey, Jerome Etienne, Jean Claude Desenclos, Christine Saura, and Didier Che. 2011. "Progress in the Surveillance and Control of Legionella Infection in France, 1998-2008." *International Journal of Infectious Diseases* 15 (1): 30–37.

CANN, K. F., D. Rh. THOMAS, R. L. SALMON, A. P. WYN-JONES, and D. KAY. 2013. "Extreme Water-Related Weather Events and Waterborne Disease." *Epidemiology and Infection* 141 (4): 671–86.

Chen, Nai-Tzu, Mu-Jean Chen, Chao-Yu Guo, Kow-Tong Chen, and Huey-Jen Su. 2014. "Precipitation Increases the Occurrence of Sporadic Legionnaires' Disease in Taiwan." *PLoS ONE* 9 (12): e114337.

Colford, John M, Ira Tager, Lawrence F Byers, Paolo Ricci, Alan Hubbard, and Robert Horner. 1999. "Methods for Assessing the Public Health Impact of Outflows from Combined Sewer Systems." *Journal of the Air & Waste Management Association* 49 (4): 454–62.

Cox, Bianca, Antonio Gasparrini, Boudewijn Catry, Frans Fierens, Jaco Vangronsveld, and Tim S. Nawrot. 2016. “Ambient Air Pollution-Related Mortality in Dairy Cattle.” *Epidemiology* 27 (6): 779–86.

Curriero, F. C., J. A. Patz, J. B. Rose, and S. Lele. 2001. “The Association between Extreme Precipitation and Waterborne Disease Outbreaks in the United States, 1948-1994.” *American Journal of Public Health* 91 (8): 1194–99.

D’Amato, G., G. Liccardi, and G. Frenguelli. 2007. “Thunderstorm-Asthma and Pollen Allergy.” *Allergy: European Journal of Allergy and Clinical Immunology* 62 (1): 11–16.

Dabrera, G., V. Murray, J. Emberlin, J. G. Ayres, C. Collier, Y. Clewlow, and P. Sachon. 2013. “Thunderstorm Asthma: An Overview of the Evidence Base and Implications for Public Health Advice.” *QJM : Monthly Journal of the Association of Physicians* 106 (3): 207–17.

Ding, Guoyong, Ying Zhang, Lu Gao, Wei Ma, Xiujun Li, Jing Liu, Qiyong Liu, and Baofa Jiang. 2013. “Quantitative Analysis of Burden of Infectious Diarrhea Associated with Floods in Northwest of Anhui Province, China: A Mixed Method Evaluation.” *PLoS ONE* 8 (6): 1–9.

dos Santos, Renato P. 2016. “Some Comments on the Reliability of NOAA’s Storm Events Database.” *arXiv Preprint*, no. 1606.06973.

Drayna, Patrick, Sandra L. McLellan, Pippa Simpson, Shun Hwa Li, and Marc H. Gorelick. 2010. “Association between Rainfall and Pediatric Emergency Department Visits for Acute Gastrointestinal Illness.” *Environmental Health Perspectives* 118 (10): 1439–43.

Dunn, Christine E., Barry Rowlingson, R. S. Bhopal, and Peter Diggle. 2012. “Meteorological Conditions and Incidence of Legionnaires’ Disease in Glasgow, Scotland: Application of Statistical Modelling,” 687–96.

Falkinham, Joseph O., Elizabeth D Hilborn, Matthew J Arduino, Amy Pruden, and Marc A Edwards. 2015. “Review Epidemiology and Ecology of Opportunistic Premise Plumbing Pathogens :” 123 (8): 749–58.

Farnham, Andrea, Lisa Alleyne, Daniel Cimini, and Sharon Balter. 2014. “Legionnaires’ Disease Incidence and Risk Factors, New York, New York, USA, 2002-2011.” *Emerging Infectious Diseases* 20 (11): 1795–1802.

Fisman, David N, Suet Lim, Gregory a Wellenius, Caroline Johnson, Phyllis Britz, Meredith Gaskins, John Maher, et al. 2005. “It’s Not the Heat, It’s the Humidity: Wet Weather Increases Legionellosis Risk in the Greater Philadelphia Metropolitan Area.” *The Journal of Infectious Diseases* 192 (12): 2066–73.

Fredricks, Gregory A., and Roger B. Nelsen. 2007. “On the Relationship between Spearman’s Rho and Kendall’s Tau for Pairs of Continuous Random Variables.” *Journal of Statistical Planning and Inference* 137 (7): 2143–50.

Gaffield, Stephen J., Robert L. Goo, Lynn A. Richards, and Richard J. Jackson. 2003. “Public Health Effects of Inadequately Managed Stormwater Runoff.” *American Journal of Public Health* 93 (9): 1527–33.

Garambois, P.A., Hélène Roux, Kévin Larnier, David Labat, and Denis Dartus. 2015. “Characterization of Catchment Behaviour and Rainfall Selection for Flash Flood Hydrological Model Calibration: Catchments of the Eastern Pyrenees.” *Hydrological Sciences Journal* 60 (3). Taylor & Francis: 424–47.

Gasparrini, Antonio. 2014. “Modeling Exposure-Lag-Response Associations with Distributed Lag Non-Linear Models.” *Statistics in Medicine* 33 (5): 881–99.

Givati, Amir, and Daniel Rosenfeld. 2004. “Quantifying Precipitation Suppression Due to Air

Pollution." *Journal of Applied Meteorology* 43 (7): 1038–56.

Groisman, Pavel Ya, Richard W. Knight, and Thomas R. Karl. 2001. "Heavy Precipitation and High Streamflow in the Contiguous United States: Trends in the Twentieth Century." *Bulletin of the American Meteorological Society* 82 (2): 219–46.

Halsby, K. D., C. a. Joseph, J. V. Lee, and P. Wilkinson. 2014. "The Relationship between Meteorological Variables and Sporadic Cases of Legionnaires' Disease in Residents of England and Wales." *Epidemiology and Infection* 142 (11): 2352–59.

Hassan, Hany M., and Mohamed A. Abdel-Aty. 2011. "Analysis of Drivers' Behavior under Reduced Visibility Conditions Using a Structural Equation Modeling Approach." *Transportation Research Part F: Traffic Psychology and Behaviour* 14 (6). Elsevier Ltd: 614–25.

Hicks, L a, C E Rose, B S Fields, M L Drees, J P Engel, P R Jenkins, B S Rouse, et al. 2007. "Increased Rainfall Is Associated with Increased Risk for Legionellosis." *Epidemiology and Infection* 135 (5): 811–17.

Ivancic, Timothy J., and Stephen B. Shaw. 2015. "Examining Why Trends in Very Heavy Precipitation Should Not Be Mistaken for Trends in Very High River Discharge." *Climatic Change* 133 (4): 681–93.

Jaccard, P. 1912. "The distribution of the flora in the alpine zone." *New phytologist*, *11*(2), 37-50.

Jonkman, Sebastiaan N., Bob Maaskant, Ezra Boyd, and Marc Lloyd Levitan. 2009. "Loss of Life Caused by the Flooding of New Orleans after Hurricane Katrina: Analysis of the Relationship between Flood Characteristics and Mortality." *Risk Analysis* 29 (5): 676–98.

Kellar, D. M M, and T. W. Schmidlin. 2012. "Vehicle-Related Flood Deaths in the United

States, 1995-2005." *Journal of Flood Risk Management* 5 (2): 153–63.

Lane, Kathryn, Kizzy Charles-Guzman, Katherine Wheeler, Zaynah Abid, Nathan Graber, and Thomas Matte. 2013. "Health Effects of Coastal Storms and Flooding in Urban Areas: A Review and Vulnerability Assessment." *Journal of Environmental and Public Health* 2013.

Lin, Cynthia, Timothy Wade, and Elizabeth Hilborn. 2015. "Flooding and Clostridium Difficile Infection: A Case-Crossover Analysis." *International Journal of Environmental Research and Public Health* 12 (6): 6948–64.

Liu, Jia Coco, Loretta J. Mickley, Melissa P. Sulprizio, Francesca Dominici, Xu Yue, Keita Ebisu, Georgiana Brooke Anderson, Rafi F A Khan, Mercedes A. Bravo, and Michelle L. Bell. 2016. "Particulate Air Pollution from Wildfires in the Western US under Climate Change." *Climatic Change* 138 (3–4). Climatic Change: 655–66.

Lowe, Dianne, Kristie L. Ebi, and Bertil Forsberg. 2013. "Factors Increasing Vulnerability to Health Effects Before, during and after Floods." *International Journal of Environmental Research and Public Health* 10 (12): 7015–67.

Mann, Andrea G, Clarence C Tam, Craig D Higgins, and Laura C Rodrigues. 2007. "The Association between Drinking Water Turbidity and Gastrointestinal Illness: A Systematic Review." *BMC Public Health* 7: 256.

Mesinger, Fedor, Geoff DiMego, Eugenia Kalnay, Kenneth Mitchell, Perry C. Shafran, Wesley Ebisuzaki, Dušan Jović, et al. 2006. "North American Regional Reanalysis." *Bulletin of the American Meteorological Society* 87 (3): 343–60.

Mitchell, Kenneth E. 2004. "The Multi-Institution North American Land Data Assimilation System (NLDAS): Utilizing Multiple GCIP Products and Partners in a Continental Distributed Hydrological Modeling System." *Journal of Geophysical Research* 109 (D7):

D07S90.

Nichols, Gordon, Chris Lane, Nima Asgari, Neville Q. Verlander, and Andre Charlett. 2009. "Rainfall and Outbreaks of Drinking Water Related Disease and in England and Wales." *Journal of Water and Health* 7 (1): 1–8.

Nixdorf-Miller, Allison, Donna M. Hunsaker, and John C. Hunsaker. 2006. "Hypothermia and Hyperthermia Medicolegal Investigation of Morbidity and Mortality from Exposure to Environmental Temperature Extremes." *Archives of Pathology and Laboratory Medicine* 130 (9): 1297–1304.

Phin, Nick, Frances Parry-Ford, Timothy Harrison, and Helen R. Stagg. 2014. "Epidemiology and Clinical Management of Legionnaires' Disease."

Pielke R.A., Jr, and M. W. Downton. 2000. "Precipitation and Damaging Floods: Trends in the United States, 1932-97." *Journal of Climate* 13 (20): 3625–37.

Pope, C. Arden, and Douglas W. Dockery. 2006. "Health Effects of Fine Particulate Air Pollution: Lines That Connect." *Journal of the Air & Waste Management Association* 56 (6): 709–42.

Robinson, Bruce, Mohammad Fahmi Alatas, Andrew Robertson, and Henry Steer. 2011. "Natural Disasters and the Lung." *Respirology* 16 (3): 386–95.

Rosenfeld, D., J. Dai, X. Yu, Z. Yao, X. Xu, X. Yang, and C. Du. 2007. "Inverse Relations Between Amounts of Air Pollution and Orographic Precipitation." *Science* 315 (5817): 1396–98.

Rowe, Scott T, and Gabriele Villarini. 2013. "Flooding Associated with Predecessor Rain Events over the Midwest United States." *Environmental Research Letters* 8 (2): 24007.

Samet, JM, SL Zeger, F Dominici, F Curriero, I Coursac, DW Dockery, J. Schwartz, and A

Zanobetti. 2000. “The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and Mortality from Air Pollution in the United States.” *Res Rep Health Eff Inst.* 94 (Pt2): 5–70.

Schwartz, J, R Levin, and R Goldstein. 2000. “Drinking Water Turbidity and Gastrointestinal Illness in the Elderly of Philadelphia.” *Journal of Epidemiology and Community Health* 54 (1): 45–51.

Setzer, Christian, and Marisa Elena Domino. 2004. “Medicaid Outpatient Utilization for Waterborne Pathogenic Illness Following Hurricane Floyd.” *Public Health Reports* 119 (5): 472–78.

Sharif, Hatim O., Md. Moazzem Hossain, Terrance Jackson, and Sazzad Bin-Shafique. 2012. “Person-Place-Time Analysis of Vehicle Fatalities Caused by fl[1] Sharif HO, Hossain MM, Jackson T, Bin-Shafique S. Person-Place-Time Analysis of Vehicle Fatalities Caused by Flash Floods in Texas. Geomatics, Nat Hazards Risk 2012;3:311–23. doi:10.1080/194.” *Geomatics, Natural Hazards and Risk* 3 (4): 311–23.

Solomon, Gina M., Mervi Hjelmroos-Koski, Miriam Rotkin-Ellman, and S. Katharine Hammond. 2006. “Airborne Mold and Endotoxin Concentrations in New Orleans, Louisiana, after Flooding, October through November 2005.” *Environmental Health Perspectives* 114 (9): 1381–86.

Soneja, Sutyajeet, Chengsheng Jiang, Jared Fisher, Crystal Romeo Upperman, Clifford Mitchell, and Amir Sapkota. 2016. “Exposure to Extreme Heat and Precipitation Events Associated with Increased Risk of Hospitalization for Asthma in Maryland, U.S.A.” *Environmental Health* 15 (1). Environmental Health: 57.

Tak, SangWoo, Bruce P. Bernard, Richard J. Driscoll, and Chad H. Dowell. 2007. “Floodwater

Exposure and the Related Health Symptoms among Firefighters in New Orleans, Louisiana 2005." *American Journal of Industrial Medicine* 50 (5): 377–82.

Thomas, Kate M, Dominique F Charron, David Waltner-Toews, Corinne Schuster, Abdel R Maarouf, and John D Holt. 2006. "A Role of High Impact Weather Events in Waterborne Disease Outbreaks in Canada, 1975 - 2001." *International Journal of Environmental Health Research* 16 (3): 167–80.

Tornevi, Andreas, Gösta Axelsson, and Bertil Forsberg. 2013. "Association between Precipitation Upstream of a Drinking Water Utility and Nurse Advice Calls Relating to Acute Gastrointestinal Illnesses." *PLoS ONE* 8 (7): 1–7.

Trenberth, Kevin E., John T. Fasullo, and Theodore G. Shepherd. 2015. "Attribution of Climate Extreme Events." *Nature Climate Change* 5 (8): 725–30.

Vanasse, Alain, Alan Cohen, Josiane Courteau, Patrick Bergeron, Roxanne Dault, Pierre Gosselin, Claudia Blais, Diane B??langer, Louis Rochette, and Fateh Chebana. 2016. "Association between Floods and Acute Cardiovascular Diseases: A Population-Based Cohort Study Using a Geographic Information System Approach." *International Journal of Environmental Research and Public Health* 13 (2): 1–12.

Wade, T. J. 2004. "Did a Severe Flood in the Midwest Cause an Increase in the Incidence of Gastrointestinal Symptoms?" *American Journal of Epidemiology* 159 (4): 398–405.

Wade, Timothy J., Cynthia J. Lin, Jyotsna S. Jagai, and Elizabeth D. Hilborn. 2014. "Flooding and Emergency Room Visits for Gastrointestinal Illness in Massachusetts: A Case-Crossover Study." *PLoS ONE* 9 (10): e110474.

Wade, Timothy J., Sukhminder K. Sandhu, Deborah Levy, Sherline Lee, Mark W. LeChevallier, Louis Katz, and John M. Colford. 2004. "Did a Severe Flood in the Midwest Cause an

Increase in the Incidence of Gastrointestinal Symptoms?" *American Journal of Epidemiology* 159 (4): 398–405.

White, Ben. 2014. "Increase in Sporadic Legionnaires ' Disease in CO Following the 2013 Floods Spring 2014 EIP MEETING."

Wilby, Robert L., Nicholas J. Clifford, Paolo De Luca, Shaun Harrigan, John K. Hillier, Richard Hodgkins, Matthew F. Johnson, et al. 2017. "The 'dirty Dozen' of Freshwater Science: Detecting Then Reconciling Hydrological Data Biases and Errors." *Wiley Interdisciplinary Reviews: Water* 4 (June): e1209.

Xiao, Xiong, Antonio Gasparrini, Jiao Huang, Qiaohong Liao, Fengfeng Liu, Fei Yin, Hongjie Yu, and Xiaosong Li. 2017. "The Exposure-Response Relationship between Temperature and Childhood Hand, Foot and Mouth Disease: A Multicity Study from Mainland China." *Environment International* 100. The Authors: 102–9.

Yang, Jun, Maigeng Zhou, Chun-Quan Ou, Peng Yin, Mengmeng Li, Shilu Tong, Antonio Gasparrini, et al. 2017. "Seasonal Variations of Temperature-Related Mortality Burden from Cardiovascular Disease and Myocardial Infarction in China." *Environmental Pollution* 224. Elsevier Ltd: 400–406.

Zanobetti, A, M P Wand, J Schwartz, and L M Ryan. 2000. "Generalized Additive Distributed Lag Models: Quantifying Mortality Displacement." *Printed in Great Britain Biostatistics* 1 (3): 279–92.

Zhang, X. C. 2005. "Spatial Downscaling of Global Climate Model Output for Site-Specific Assessment of Crop Production and Soil Erosion." *Agricultural and Forest Meteorology* 135 (1–4): 215–29.

Zhao, Wei, and Ainong Li. 2015. "A Review on Land Surface Processes Modelling over

Complex Terrain.” *Advances in Meteorology* 2015.

APPENDIX A

THE "COUNTYWEATHER" R PACKAGE

As part of this research, we published the "countyweather" package on the Comprehensive R Archive Network in October 2016; a development version currently exists online on GitHub. All of the package's code is open source. This appendix describes this software package and is included as a tutorial in the published package.

While data from weather stations is available at the specific location of the weather station, it is often useful to have estimates of daily or hourly weather aggregated on a larger spatial level. For U.S.-based studies, it can be particularly useful to be able to pull time series of weather by county. For example, the health data used in environmental epidemiology studies is often aggregated at the county level for U.S. studies, making it very useful for environmental epidemiology applications to be able to create weather datasets by county.

This package builds on functions from the rnoaa package to identify weather stations within a county based on its FIPS code and then pull weather data for a specified date range from those weather stations. It then does some additional cleaning and aggregating to produce a single, county-level weather dataset. Further, it maps the weather stations used for that county and date range and allows you to create and write datasets for many different counties using a single function call.

If you are pulling weather data from single weather station, you should use rnoaa directly. However, countyweather allows you to pull and aggregate data from weather stations more easily at the county level for the US.

Required set-up for this package

To use this package, you will need an API key from NOAA to be able to access the weather data. This API key is input with some of your data requests to NOAA within functions in this package. You can request an API key from NOAA here: http://www.ncdc.noaa.gov/cdo-web/token. You should keep this key private.

Once you have this NOAA API key, you'll need to pass it through to some of the functions in this package that pull data from NOAA. The most secure way to use this API key is to store it in your .Renviron configuration file. Then you can save it as the value of an object in R code or R markdown documents without having to include the key itself in the script. To store the NOAA API key in your .Renviron configuration file, first check and see if you already have an .Renviron file in your home directory. You can check this by running the following from your R command line:

```
any(grepl("^\\.Renviron", list.files("~", all.files = TRUE)))
```

If this call returns TRUE, then you already have an .Renviron file.

If you already have it, open that file (for example, with system("open ~/.Renviron")). If you do not yet have an .Renviron file, open a new text file (in RStudio, do this by navigating

to *File > New File > Text File*) and save this text file as .Renviron in your home directory. If prompted with a complaint, you DO want to use a filename that begins with a dot .

Once you have opened or created an .Renviron file, type the following into the file, replacing "your_emailed_key" with the actual string that NOAA emails you:

```
noaakey=your_emailed_key
```

Do not put quotation marks or anything else around your key. Do make sure to add a blank line as the last line of the file. If you find you're having problems getting this to work, go back and confirm that you've included a blank line as the last line in your .Renviron file. This is the most common reason for this part not working.

Next, you'll need to restart R. Once you restart R, you can get the value of this NOAA API key from .Renviron anytime with the call Sys.getenv("noaakey"). Before using functions that require the API key, set up the object rnoaakey to have your NOAA API key by running:

```
options("noaakey" = Sys.getenv("noaakey"))
```

This will pull your NOAA API key from the .Renviron file and save it as the object noaakey, which functions in this package need to pull weather data from NOAA's web services. You will want to put this line of code as one of the first lines of code in any R script or R Markdown file you write that uses functions from this package.

Basic examples of using the package

Weather data is collected at weather station, and there are often multiple weather stations within a county. The countyweather package allows you to pull weather data from all stations in a specified county over a specified date range. The two main functions in the countyweather package are daily_fips and hourly_fips, which pull daily and hourly weather data, respectively. By default, the weather data pulled from all weather stations in a county will then be averaged for each time point to create an average time series of daily or hourly measurements for that county. There is also an option that allows the user to opt out of the default aggregation across weather stations, and instead pull separate time series for each weather station in the county. This option is explained in more detail later in this document. Opting out of the default aggregation can be useful if you would like to use a method other than a simple average to aggregate across weather stations within a county.

Throughout, functions in this package identify a county using the county's Federal Information Processing Standard (FIPS) code. FIPS codes are 5-digit codes that uniquely identify every U.S. county. The first two digits of a county FIPS code specify state and the last three specify the county within the state. This package pulls data based on FIPS designations as of the 2010 Census. Users will not be able to pull data for the few FIPS codes that have undergone substantial changes since 2010 - for a list of those codes see the Census Bureau's summary of these counties for the 2010s.

Currently, this package can pull daily and hourly weather data for variables like temperature and precipitation. For resources with complete lists of weather variables available through this package, as well as sources of this weather data, see the section later in this document titled "More on the weather data".

Pulling daily data

The daily_fips function can be used to pull daily weather data for all weather stations within the geographic boundaries of a county. This daily weather data comes from NOAA's Global Historical Climatology Network. When pulling data for a county, the user can specify date ranges (date_min, date_max), which weather variables to include in the output dataset (var), and restrictions on how much non-missing data a weather station must have over the time period to be included when generating daily county average values (coverage). This function will pull any available data for weather stations in the county under the specified restrictions and output both a dataset of average daily observations across all county weather stations, as well as a map plotting the stations used in the county-wide averaged data.

Here is an example of creating a dataset with daily precipitation for Miami-Dade county (FIPS code = 12086) for August 1992, when Hurricane Andrew stuck:

```
andrew_precip <- daily_fips(fips = "12086", date_min = "1992-08-01",
                            date_max = "1992-08-31", var = "prcp")
names(andrew_precip)
## [1] "daily_data"       "station_metadata" "station_map"
```

The output from this function call is a list that includes three elements: a daily time series of weather data for the county (andrew_precip$daily_data), a dataframe with meta-data about the weather stations used to create the time series data, as well as statistical information about the weather values pulled from these stations (andrew_precip$station_metadata), and a map showing the locations of weather stations included in the county-averaged dataset (andrew_precip$station_map).

Here are the first few rows of the daily_data dataset:

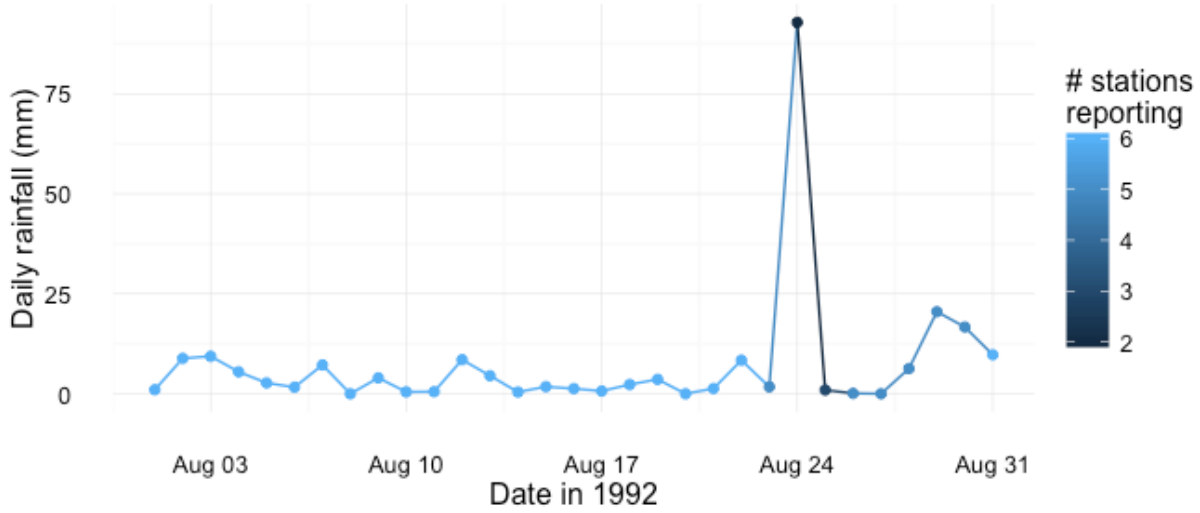
```
head(andrew_precip$daily_data)
## # A tibble: 6 × 3
##         date     prcp prcp_reporting
##       <date>    <dbl>          <int>
## 1 1992-08-01 1.016667              6
## 2 1992-08-02 8.850000              6
## 3 1992-08-03 9.366667              6
## 4 1992-08-04 5.483333              6
## 5 1992-08-05 2.716667              6
## 6 1992-08-06 1.633333              6
```

The dataset includes columns for date (date), precipitation (in mm, prcp), and also the number of stations used to calculate each daily average precipitation observation (prcp_reporting).

This function performs some simple data cleaning and quality control on the weather data originally pulled from NOAA's web services; see the "More on the weather data" section later in this document for more details, including the units for the weather observations collected by this function.

Here is a plot of this data, with colors used to show the number of stations included in each daily observation:

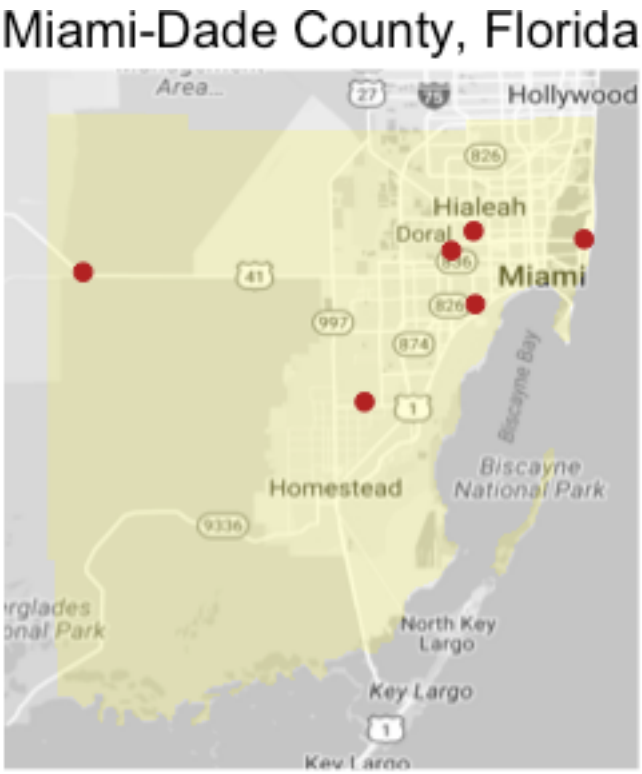
```
library(ggplot2)
ggplot(andrew_precip$daily_data, aes(x = date, y = prcp, color = prcp_reporting)) +
  geom_line() + geom_point() + theme_minimal() +
  xlab("Date in 1992") + ylab("Daily rainfall (mm)") +
  scale_color_continuous(name = "# stations\nreporting")
```



From this plot, you can see both the extreme precipitation associated with Hurricane Andrew (Aug. 24) and that the storm knocked out quite a few of the weather stations normally available.

A map is also included in the output of `daily_fips` with the stations used for the county average, as the `station_map` element:

```
andrew_precip$station_map
```



This map uses U.S. Census TIGER/Line shapefiles (vintage 2011) and functions from the ggmap package to overlay weather station locations on a shaped map showing the county's boundaries.

The station_metadata dataframe gives information about all of the stations contributing data to the daily_datadataframe, as well as information about how the values by each station vary within each weather variable. If a weather station is contributing data for multiple variables, it will show up in this dataframe multiple times. Here's what the station_metadata dataframe looks like for the andrew_precip list:

```
andrew_precip$station_metadata
```

```
##            id                                      name   var latitude
##         <chr>                                     <chr> <chr>    <dbl>
## 1 USC00083909                          HIALEAH, FL US  prcp 25.81750
## 2 USC00087020                      PERRINE 4 W, FL US  prcp 25.58190
## 3 USC00088780         TAMIAMI TRAIL 40 MI. BEND, FL US  prcp 25.76080
## 4 USW00012839       MIAMI INTERNATIONAL AIRPORT, FL US  prcp 25.79050
## 5 USW00012859 MIAMI WEATHER SERVICE OFFICE CITY, FL US  prcp 25.71667
## 6 USW00092811                      MIAMI BEACH, FL US  prcp 25.80630
## # ... with 6 more variables: longitude <dbl>, calc_coverage <dbl>,
## #   standard_dev <dbl>, min <dbl>, max <dbl>, range <dbl>
```

For each station, the dataframe gives an id and name, as well as latitude and longitude. var indicates the variable for which the station is pulling data. If a

station is contributing data for multiple variables, that station will show up in the dataframe once for each of those variables. For each variable and station combination, the dataframe also shows calc_coverage, which is the calculated percent of non-missing values. You can filter these by using the daily_fips option coverage. standard_dev gives the standard deviation for each sample of weather data from each station and weather variable, min and max give the minimum and maximum values, and range gives the range of these values. These last four statistical calculations (standard deviation, maximum, minimum, and range) are only included for the seven core hourly weather variables (which include wind_direction, wind_speed, ceiling_height, visibility_distance, temperature, and temperature_dewpoint – for more details on these variables, see the "More on the weather data" section below). The values of these columns are set to "NA" for other variables, such as quality flag data.

If you are interested in looking at the weather values for certain stations, you can use the average_data = FALSE option in daily_fips. For more on this option and a few others, see the "Further options available in the package" section below.

Pulling hourly data

You can use the hourly_fips function to pull hourly weather data by county from NOAA's Integrated Surface Data (ISD) weather dataset. In this case, NOAA's web services will not identify weather stations by FIPS, so instead this function will pull all stations within a certain radius of the county's population mean center to represent weather within that county. While there are seven main weather variables that are possible to pull (listed below in the "More on the weather data" section), temperature and wind_speed tend to be non-missing most often.

An estimated radius is calculated for each county using 2010 U.S. Census Land Area data – each county is assumed to be roughly circular. The calculated radius (in km), as well as the longitude and latitude of the geographic center for each county are included as elements in the list returned from hourly_fips.

Here is an example of pulling hourly data for Miami-Dade, for the year of Hurricane Andrew. While daily weather data can be pulled using a date range specified to the day, hourly data can only be pulled by year (for one or multiple years) using the year argument:

```
andrew_hourly <- hourly_fips(fips = "12086", year = 1992,
                             var = c("wind_speed", "temperature"))
names(andrew_hourly)
## [1] "hourly_data"      "station_metadata" "station_map"
## [4] "radius"           "lat_center"       "lon_center"
```

The output from this call is a list object that includes six elements. andrew_hourly$hourly_data is an hourly time series of weather data for the county. The other five elements, station_metadata, station_map, radius, lat_center, and lon_center, are explained in more detail below.

Here are the first few rows of the hourly_data dataset:

```
head(andrew_hourly$hourly_data)
```

```
## # A tibble: 6 × 5
##             date_time temperature wind_speed temperature_reporting
##                 <dttm>       <dbl>      <dbl>                 <int>
## 1 1992-01-01 00:00:00    19.63333   2.725000                     3
## 2 1992-01-01 01:00:00    19.43333   2.450000                     3
## 3 1992-01-01 02:00:00    19.03333   2.975000                     3
## 4 1992-01-01 03:00:00    19.03333   2.450000                     3
## 5 1992-01-01 04:00:00    18.53333   2.200000                     3
## 6 1992-01-01 05:00:00    18.50000   2.233333                     3
## # ... with 1 more variables: wind_speed_reporting <int>
```

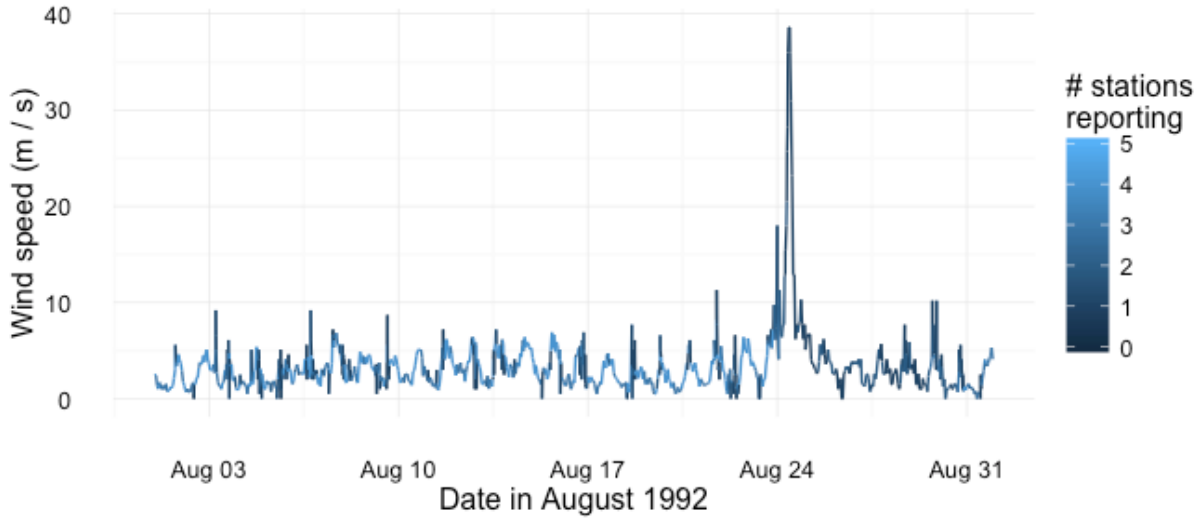
If you need to get the timestamp for each observation in local time, you can use the add_local_time function from the countytimezones package to do that:

```
andrew_hourly_data <- as.data.frame(andrew_hourly$hourly_data)


library(countytimezones)
andrew_hourly_data <- add_local_time(df = andrew_hourly_data, fips = "12086",
                                     datetime_colname = "date_time")
head(andrew_hourly_data)
##             date_time temperature wind_speed temperature_reporting
## 1 1992-01-01 00:00:00    20.00000      2.600                     4
## 2 1992-01-01 01:00:00    19.85000      1.960                     4
## 3 1992-01-01 02:00:00    19.03333      2.975                     3
## 4 1992-01-01 03:00:00    19.42500      1.960                     4
## 5 1992-01-01 04:00:00    18.53333      1.760                     3
## 6 1992-01-01 05:00:00    18.60000      2.575                     4
##   wind_speed_reporting       local_time local_date         local_tz
## 1                    5 1991-12-31 19:00 1991-12-31 America/New_York
## 2                    5 1991-12-31 20:00 1991-12-31 America/New_York
## 3                    4 1991-12-31 21:00 1991-12-31 America/New_York
## 4                    5 1991-12-31 22:00 1991-12-31 America/New_York
## 5                    5 1991-12-31 23:00 1991-12-31 America/New_York
## 6                    4 1992-01-01 00:00 1992-01-01 America/New_York
```

Here is a plot of hourly wind speeds for Miami-Dade County, FL, for the month of Hurricane Andrew:

```
library(dplyr)
library(lubridate)
```

```
to_plot <- andrew_hourly$hourly_data %>%
  filter(months(date_time) == "August")
ggplot(to_plot, aes(x = date_time, y = wind_speed,
                    color = wind_speed_reporting)) +
  geom_line() + theme_minimal() +
  xlab("Date in August 1992") +
  ylab("Wind speed (m / s)") +
  scale_color_continuous(name = "# stations\nreporting")
```



Again, the intensity of conditions during Hurricane Andrew is clear, as is the reduction in the number of reporting stations during the storm.

The list object returned by `hourly_fips` also includes a map of station locations (`station_map`):

```
andrew_hourly$station_map
```

Because hourly data is pulled by radius from each county's geographic center, this plot includes the calculated radius from which stations are pulled. This radius is calculated for each county using 2010 U.S. Census Land Area data. U.S. Census TIGER/Line shapefiles are used to provide county outlines, included on this plot as well. Because stations are pulled within a radius from the county's center, stations from outside of the county's boundaries may sometimes be providing data for that county.

Other list elements returned by `hourly_fips` include `station_metadata`, `radius`, `lat_center`, and `lon_center`. `Radius` is the estimated radius (in km) for the county calculated using 2010 U.S. Census Land Area data – the county is assumed to be roughly circular. `lat_center` and `lon_center` are the longitude and latitude of the geographic center for the county, respectively.

The `station_metadata` dataframe gives information about all of the stations contributing data to the `hourly_data` dataframe, as well as information about how the values by each station vary within each weather variable. If a weather station is contributing data for multiple variables, it will show up in this dataframe multiple times. Here's what the `station_metadata` dataframe looks like for the `andrew_hourly` list:

```
andrew_hourly$station_metadata
```

```
## # A tibble: 8 × 15
```

```
##     usaf  wban      station                station_name         var
##    <chr> <chr>        <chr>                       <chr>       <chr>
## 1 722029  <NA>    722029-NA        KENDALL TAMIAMI EXEC  wind_speed
## 2 722026 12826 722026-12826       HOMESTEAD AFB AIRPORT  wind_speed
```

```
## 3 722020 12839 722020-12839 MIAMI INTERNATIONAL AIRPORT  wind_speed
## 4 722024  <NA>    722024-NA                   OPA LOCKA  wind_speed
## 5 722029  <NA>    722029-NA        KENDALL TAMIAMI EXEC temperature
## 6 722026 12826 722026-12826       HOMESTEAD AFB AIRPORT temperature
## 7 722020 12839 722020-12839 MIAMI INTERNATIONAL AIRPORT temperature
## 8 722024  <NA>    722024-NA                   OPA LOCKA temperature
## # ... with 10 more variables: calc_coverage <dbl>, standard_dev <dbl>,
## #   range <dbl>, ctry <chr>, state <chr>, elev_m <dbl>, begin <dbl>,
## #   end <dbl>, longitude <dbl>, latitude <dbl>
```

`usaf` and `wban` are station ids. `station` is a unique identifier for each station – usaf and wban ids have been pasted together, separated by "-". (Note: values for `wban` or `usaf` are sometimes missing (originally indicated by "99999" or "999999"), which could result in a `station` value like `722024-NA`.) `station_name` is the name for each station, and `var` indicates the variable for which the station is pulling data. If a station is contributing data for multiple variables, that station will show up in the dataframe once for each of those variables. For each variable and station combination, the dataframe also shows `calc_coverage`, which is the calculated percent of non-missing values. You can filter these by using the `hourly_fips` option `coverage`. `standard_dev` gives the standard deviation for each sample of weather data from each station and weather variable, and `range` gives the range of these values. Here, we can see in row 7 that the OPA LOCKA station has a very low percent coverage for temperature (0.0018), and a correspondingly high standard deviation (17.94). If you are interested in looking at the weather values for certain stations, you can use the `average_data = FALSE` option in `hourly_fips`. For more on this option and a few others, see the "Further options available in the package" section below. The dataframe also gives station countries, states, elevation (in meters), the earliest and latest dates for which the station has available data (`begin` and `end`, respectively), longitude, and latitude.

Writing out time series files

There are a few functions that allow the user to write out daily or hourly time series datasets for many different counties to a specified local directory, as well as plots of this data. For daily weather data, see the functions `write_daily_timeseries` and `plot_daily_timeseries`. For hourly, see `write_hourly_timeseries` and `plot_hourly_timeseries`.

For example, if we wanted to compare daily weather in the month of August for three counties in southern Florida, we could run:

```
fl_counties <- c("12086", "12087", "12011")


write_daily_timeseries(fips = fl_counties, date_min = "1992-08-01",
                  date_max = "1992-08-31", var = "prcp",
                  out_directory = "~/Documents/andrew_data")
```

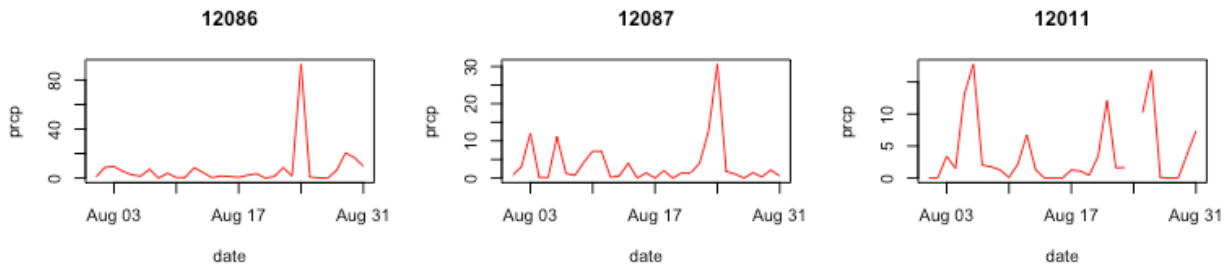
The write_daily_timeseries function saves each county's time series as a separate file in a subdirectory called "data" of the directory specified in the out_directory option.
The data_type argument allows the user to specify either .rds or .csv files (the default is to write .rds files). Each file is a time series dataframe of daily weather data. A dataframe of station

metadata is saved in a second subdirectory called "metadata", and maps showing locations of weather stations contributing to each time series are saved in a subdirectory called "maps." At this stage, if you were to include a county in the fips argument without available data, a file would not be created for that county.

The function plot_daily_timeseries creates and saves plots for each of these files. (Note: the data_type argument for this function also defaults to read .rds files, so if you chose to write .csv files, make sure to change that argument in this function as well to data_type = "csv".)

```
plot_daily_timeseries("prcp", data_directory = "~/Documents/andrew_data/data",
                      plot_directory = "~/Documents/andrew_data/plots",
                      date_min = "1992-08-01", date_max = "1992-08-31")
```

Here's an example of what the time series plots for the three Florida counties would look like:



Further options available in the package

coverage

For hourly_fips, daily_fips, and time series functions, the user can choose to filter out any stations that report variables for less than a certain percent of the specified date range (coverage). For example, if you were to set coverage to 0.90, only stations that reported non-missing values at least 90% of the time over the specified date range would be included in your data.

average_data

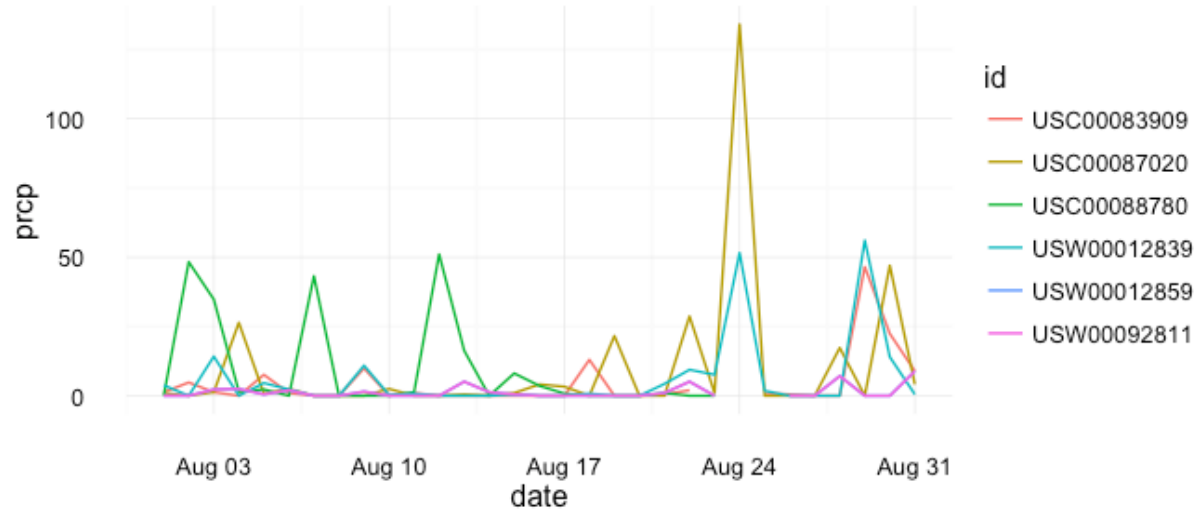
In both daily_fips and hourly_fips, the default is to return a single daily average for the county for each day in the time series, giving the value averaged across all available stations on that day. However, there is also an option called average_data which allows the user to specify whether they would like the weather data returned before it has been averaged across stations. If this argument is set to FALSE, the functions will return separate daily data for each station in the county. For our Hurricane Andrew example, we can specify average_data = FALSE:

```
not_averaged <- daily_fips(fips = "12086",
                           date_min = "1992-08-01",
                           date_max = "1992-08-31",
                           var = "prcp", average_data = FALSE,
                           station_label = TRUE)
```

```
not_averaged_data <- not_averaged$daily_data

head(not_averaged_data)

## # A tibble: 6 × 3

##            id       date  prcp

##         <chr>     <date> <dbl>

## 1 USC00083909 1992-08-01   1.3

## 2 USC00083909 1992-08-02   4.8

## 3 USC00083909 1992-08-03   1.3

## 4 USC00083909 1992-08-04   0.0

## 5 USC00083909 1992-08-05   7.6

## 6 USC00083909 1992-08-06   1.0

unique(not_averaged_data$id)

## [1] "USC00083909" "USC00087020" "USC00088780" "USW00012839" "USW00012859"

## [6] "USW00092811"
```

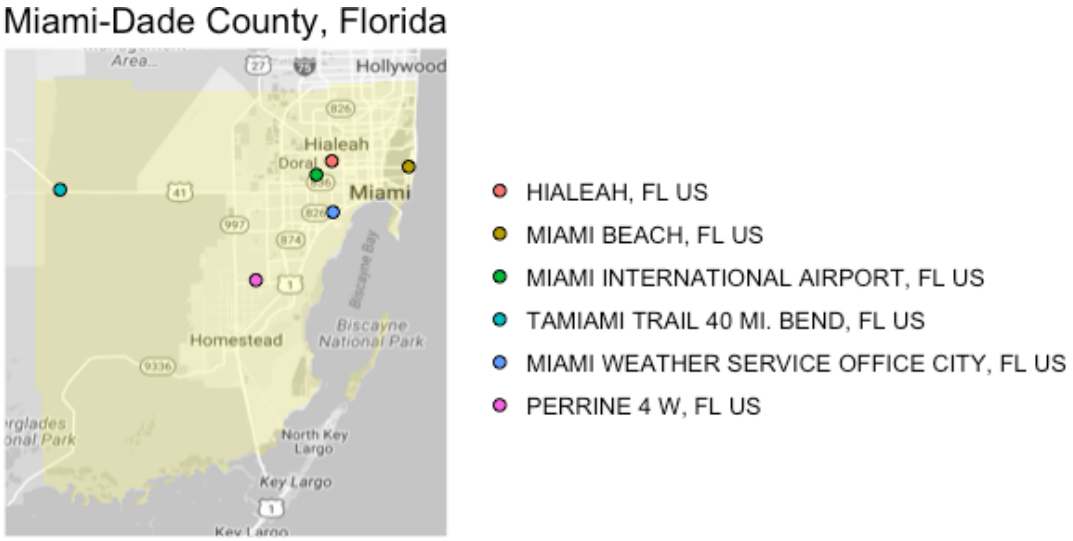
In this example, there are six stations contributing weather data to the time series. We can plot the data by station to get a sense for how values from each station compare, and which stations were presumably knocked out by the storm, with different colors used to show values for different stations:

```
library(ggplot2)

ggplot(not_averaged_data, aes(x = date, y = prcp,

                              colour = id)) +

  geom_line() +

  theme_minimal()
```



It might be interesting here to compare this plot with the station map, this time with station labels included (done using station_label = TRUE when we pulled this data using daily_fips):

```
not_averaged$station_map
```



Quality Flags

The hourly Integrated Surface Data includes quality codes for each of the main weather variables. For more information about the hourly weather variables, see the "More on the weather data" section below. We can use these codes to remove suspect or erroneous values from our data. The values in `wind_speed_quality`, for example, take on the following values: (Values in this table were pulled from the ISD documentation file.)

code	definition
0	Passed gross limits check
1	Passed all quality control checks
2	Suspect
3	Erroneous
4	Passed gross limits check, data originate from an NCEI data source

code	definition
5	Passed all quality control checks, data originate from an NCEI data source
6	Suspect, data originate from an NCEI data source
7	Erroneous, data originate from an NCEI data source
9	Passed gross limits check if element is present

Because it doesn't make sense to average these codes across stations, the codes should only be pulled when using the option to pull station-specific values (average_data = FALSE).

```
ex <- hourly_fips("12086", 1992, var = c("wind_speed", "wind_speed_quality"),
                  average_data = FALSE)
ex_data <- ex$hourly_data
head(ex_data)
## # A tibble: 6 × 7
##   usaf_station wban_station           date_time latitude longitude
##          <dbl>        <dbl>              <dttm>    <dbl>     <dbl>
## 1       722029           NA 1992-01-01 00:00:00    25.65   -80.433
## 2       722029           NA 1992-01-01 01:00:00    25.65   -80.433
## 3       722029           NA 1992-01-01 02:00:00    25.65   -80.433
## 4       722029           NA 1992-01-01 03:00:00    25.65   -80.433
## 5       722029           NA 1992-01-01 04:00:00    25.65   -80.433
## 6       722029           NA 1992-01-01 05:00:00    25.65   -80.433
## # ... with 2 more variables: wind_speed <dbl>, wind_speed_quality <chr>
```

We can replace all wind speed observations with quality codes of 2, 3, 6, or 7 with NAs.

```
ex_data$wind_speed_quality <- as.numeric(ex_data$wind_speed_quality)
ex_data$wind_speed[ex_data$wind_speed_quality %in% c(2, 3, 6, 7)] <- NA
```

More on the weather data

Daily weather data

Functions in this package that pull daily weather values (daily_fips(), for example) are pulling data from the Daily Global Historical Climatology Network (GHCN-Daily) through NOAA's FTP server. The data is archived at the National Centers for Environmental Information (NCEI) (formerly the National Climatic Data Center (NCDC)), and spans from the 1800s to the current year.

Users can specify which weather variables they would like to pull. The five core daily weather variables are precipitation (prcp), snowfall (snow), snow depth (snwd), maximum temperature (tmax) and minimum temperature (tmin). The daily weather data is filtered so that included weather variables fall within a range of possible values. These ranges were chosen to include national maximum recorded values.

Variable	Description	Units	Most extreme value
prcp	precipitation	mm	1100 mm
snow	snowfall	mm	1600 mm
snwd	snow depth	mm	11500 mm
tmax	maximum temperature	degrees Celsius	57 degrees C
tmin	minumum temperature	degrees Celsius	-62 degrees C

tmax, tmin, and prcp were originally recorded in tenths of units, and are listed as such in NOAA documentation. These values are converted to standard units (degrees Celsius and mm, respectively) in countyweather output.

There are several additional, non-core variables available. For example, acmc gives the "average cloudiness midnight to midnight from 30-second ceilometer data (percent)." The complete list of available weather variables can be found under 'element' from the GHCND's readme file.

While the datasets resulting from functions in this package return a cleaned and aggregated dataset, Menne et al. (2012) give more information about the raw data in the GHCND database.

Hourly weather data

Hourly weather data in this package is pulled from NOAA's Integrated Surface Data (ISD), and is available from 1901 to the current year. The data is archived at the National Centers for

Environmental Information (NCEI) (formerly the National Climatic Data Center (NCDC)), and is also pulled through NOAA's FTP server.

The seven core hourly weather variables are wind_direction, wind_speed, ceiling_height, visibility_distance, temperature, temperature_dewpoint, and air_pressure. Values in this table were pulled from the ISD documentation file.

Variable	Description	Units	Minimum	Maximum
wind_direction	The angle, measured in a clockwise direction, between true north and the direction from which the wind is blowing	Angular Degrees	1	360
wind_speed	The rate of horizontal travel of air past a fixed point	Meters per Second	0	90
ceiling_height	The height above ground level of the lowest cloud or obscuring phenomena layer aloft with 5/8 or more summation total sky cover, which may be predominately opaque, or the vertical visibility into a surface-based obstruction	Meters	0	22000 (indicates 'Unlimited')
visibility_distance	The horizontal distance at which an object can be seen and identified	Meters	0	160000
temperature	The temperature of the air	Degrees Celsius	-93.2	61.8
temperature_dewpoint	The temperature to which a given parcel of air must be cooled at constant pressure	Degrees Celsius	-98.2	36.8

Variable	Description	Units	Minimum	Maximum
	and water vapor content in order for saturation to occur			
air_pressure	The air pressure relative to Mean Sea Level	Hectopascals	860	1090

There are other columns available in addition to these weather variables, such as quality codes (e.g., wind_direction_quality — each of the main weather variables has a corresponding quality code that can be pulled by adding _quality to the end of the variable name).

For more information about the weather variables described in the above table and other available columns, see the ISD documentation file.

Error and warning messages you may get

Not able to pull data from a station

The following error message will come up after running functions pulling daily data if there isn't available data (for your specified date range, coverage, and weather variables) for a particular station or stations:

In rnoaa::meteo_pull_monitors(monitors = stations, keep_flags = FALSE,: The following stations could not be pulled from the GHCN ftp: USR0000FTEN Any other monitors were successfully pulled from GHCN.

The following error message will come up after running functions pulling hourly data (hourly_fips()) if there isn't available data for any of the stations in your specified county. Note: some weather variables tend to be missing more often than others.

```
Error in isd_monitors_data(fips = fips, year = x, var = var, radius = radius): None of the stations had available data.
```

The following error message will come up after running write_daily_timeseries or write_hourly_timeseries if the function is unable to pull data for a particular fips code in your fips vector:

```
Unable to pull weather data for FIPS code "(specified fips code)" for the specified percent coverage, year(s), and/or weather variables.
```

Need an API key for NOAA data

If you run functions that use NOAA API calls without first requesting an API key from NOAA and setting up the key in your R session, you will see the following error message:

```
Error in getOption("noaakey", stop("need an API key for NOAA data")) :
```

```
  need an API key for NOAA data
```

You might also see this warning message:

```
Warning message:

Error: (400) - Token parameter is required.
```

If you get one of these messages, run the code:

```
options("noaakey" = Sys.getenv("noaakey"))
```

and then try again. If you still get an error, you may not have set up your NOAA API key correctly in your .Renviron file. See the "Required set-up" section of this document for more details on doing that correctly.

NOAA web services down

Sometimes, some of NOAA's web services will be off-line. In this case, you may get an error message when you try to pull data like:

```
Error in gzfile(file, mode) : cannot open the connection
```

or

```
Error in tt$results : $ operator is invalid for atomic vectors In addition: Warning message:

Error: (500) - An error occurred while servicing your request.
```

In this case, re-starting your R session might fix the problem. If not, wait a few hours and then try again.

Other errors

If you get other error messages or run into problems with this package, please submit a reproducible example on this repository's Issues page.

References

Menne, Matthew J, Imke Durre, Russell S Vose, Byron E Gleason, and Tamara G Houston. 2012. "An Overview of the Global Historical Climatology Network-Daily Database." *Journal of Atmospheric and Oceanic Technology* 29 (7): 897–910.